Title:

Global soil pollution by toxic metals threatens agriculture and human health

Authors:

Deyi Hou[1*,#], Xiyue Jia[1,#], Liuwei Wang[1], Steve P. McGrath[2], Yong-Guan Zhu[3,4], Qing Hu[5], Fang-Jie Zhao[6], Michael S. Bank[7, 8], David O'Connor [9], Jerome Nriagu[10]

Affiliations:

[1]Tsinghua University, School of Environment; Beijing, China.

[2]Rothamsted Research, Sustainable Soils and Crops; Harpenden, United Kingdom.

[3]Chinese Academy of Sciences, Research Center for Eco-Environmental Sciences; Beijing, China.

[4]Chinese Academy of Sciences, Institute of Urban Environment; Xiamen, China.

[5]Southern University of Science and Technology, Engineering Innovation Centre (Beijing); Shenzhen, China.

[6]Nanjing Agricultural University, College of Resources and Environmental Sciences; Nanjing, China.

[7]Institute of Marine Research, Bergen; Norway.

[8]University of Massachusetts Amherst, Department of Environmental Conservation, Amherst, MA, USA

[9]Royal Agricultural University, School of Real Estate & Land Management; Cirencester, UK.

[10]University of Michigan, School of Public Health; Ann Arbor, USA.

*Corresponding author. Email: houdeyi@tsinghua.edu.cn

#These authors contributed equally

Abstract:

Toxic metal pollution is ubiquitous in soils, yet its worldwide distribution is unknown. Here we analyze a global database of soil pollution by arsenic, cadmium, cobalt, chromium, copper, nickel, and lead at 796,084 sampling points from 1493 regional studies and used machine learning techniques to map areas with exceedance of agricultural and human health thresholds. We reveal a previously unrecognized high risk, metal-enriched zone in low-latitude Eurasia, which is attributed to influential climatic, topographic, and anthropogenic conditions. This feature can be regarded as a signpost for the Anthropocene era. We show that 14% to 17% of cropland is affected by toxic metal pollution globally and estimate that between 0.9 and 1.4 billion people live in regions of heightened public health and ecological risks.

One-Sentence Summary:

Global soil pollution with toxic metals and the potential impacts on agriculture and human health are analyzed using machine learning techniques.

Main Text:

Soil provides the basis for nearly 95% of food consumed by human beings (*1*). As the human population continues to grow and living standards improve, global food production needs to increase by 35% to 56% by 2050 (*2*). This puts substantial pressure on non-renewable soil resources, the degradation of which already threatens the livelihoods of 1.3 billion people globally (*3*). The UN Food and Agriculture Organization (FAO) warns that 90% of global soil resources may be at risk by 2050, due to soil erosion, excessive usage of fertilizers and pesticides, and industrial pollution (*4, 5*). Often overlooked in the matter of soil quality is soil pollution by toxic heavy metals and metalloids (herein toxic metals), which reduces crop yields and results in unsafe food. While some metals like cobalt (Co) and copper (Cu) are essential in small amounts for biological functioning, their bioaccumulation in organisms, including crops, can render them toxic in the human food chain. Furthermore, toxic metals are non-degradable, and therefore accumulate over decadal time scales in soils (*6*-*8*).

Global soil pollution by toxic metals has been studied for decades (*9*); however, quantitative estimates of their impact on soil quality and spatially explicit mapping of soil pollution on a global scale are lacking. A few regional and country-scale investigations have provided concerning data on this issue. For instance, a national survey in China found that 19% of agricultural soils exceeded soil quality standards, with arsenic (As, a metalloid), cadmium (Cd), Cu, and nickel (Ni) accounting for the majority of exceedances (*10*). A study on toxic metals across 27 European countries showed that 28% of soils exceeded thresholds (*11*).

There are two main sources of toxic metals in soil: geogenic and anthropogenic. Toxic metals are ubiquitous in bedrocks, the natural soil parent materials, and occur in varying concentrations. Some types of parent rock (e.g. basalt, shale) as well as primary minerals (e.g. pyrite, sphalerite) contain elevated levels of As, Cd, Cu, and Ni, due to the high affinity of sulfur for these metals (*8, 12*). During the geologic weathering and soil-forming processes, toxic metals are continuously released from soil parent materials (*13, 14*). Some toxic metals may also be transported in the atmosphere following volcanic emissions and wind erosion and subsequently deposited in surface soil *(13, 15)*. Due to translocation and transformation mechanisms during pedogenesis, toxic metals may accumulate in soil due to fixation in crystal lattices, binding with clay minerals via electrostatic forces, or complexation with organic matter and iron (Fe) oxyhydroxides, which can lead to high natural background of toxic metal concentrations in certain soil environments (*12*, *16*, *17*).

Anthropogenic sources of toxic metals in pedosphere include agricultural, household, and industrial activities. Significant metal contamination of soils commenced at the beginning of the Anthropocene (e.g. Bronze age), particularly as a result of metal mining and processing (*13, 18*). Mining activities transfer huge quantities of rock, often with high metal concentrations from the underground to the surface. This leads to soil pollution by leachate and runoff from mining waste, irrigation of cropland with polluted water, wind-eroded waste rocks, and atmospheric deposition originating from metal smelters (*6*, *19*). Metal pollution at a given location may be transported across long distances as evidenced by ice cores recovered from Greenland, which reveal that intensive mining and smelting activities in the Greek and Roman times caused pronounced pulse in metal contamination at hemispheric scales (*20*). Elevated toxic metal contents are also embedded in industrial infrastructure (machinery, bridges, transport systems, cables, to buildings), and agricultural and household products (such as phosphorus fertilizers, paints, and batteries), which can contribute significantly to the toxic metal burden in soil ecosystems (*21*).

The spatial distribution of toxic metals in soil depends on a dynamic and complex balance between input and output processes. The main output pathways include leaching, soil erosion by surface runoff, plant uptake and crop harvest (*13, 17, 22*). Redistribution of toxic metals may occur in the vertical dimension of soil profiles due to soil-plant interactions. The plant-pump effect, for instance, transports toxic metals from deeper soil (e.g. C horizons) to surficial soil (e.g. O horizons), where they accumulate (*17*). Toxic metals in soil may also migrate at regional scales due to biovolatization, wind-borne soil suspension, forest fires, and other perturbances (*13*, *15*). Based on these migration mechanisms, it has been suggested that certain environmental and socioeconomic factors, including topography, climate, soil texture, and human activities may be used as predictors to evaluate toxic metal distribution across large spatial scales (*11*, *23*-*25*).

The combination of recent developments in machine learning technologies and the availability of expansive measurement data now make it possible to undertake a systematic assessment of global soil pollution for seven toxic metals: As, Cd, Co, chromium (Cr), Cu, Ni, and lead (Pb). We hypothesized that soil pollution, on a global scale, would be governed by both direct and indirect effects of biogeophysical and anthropogenic factors. Using machine learning models, we identified and analyzed multi-layered and non-linear relationships, and developed a robust and spatially explicit, continuous prediction of toxic metal exceedances based on sparsely distributed global data.

Global toxic metal exceedances

We have compiled 796,084 datapoints of soil concentrations of the key toxic metals from 1493 regional studies covering diverse climate zones, geologic settings, and land use types (figs. S1 and S2) (*26*). Data quality assurance procedures were followed to ensure that the data were reliable and representative of regional metal concentrations, and appropriate analytical methods were used to ensure robust measurements (*26*). Samples collected from studies focusing on contaminated sites were excluded to avoid bias toward highly enriched localized areas. Soil concentrations in 10 km by 10 km pixels were converted to binary data using a set of agricultural thresholds (AT) and human health and ecological thresholds (HHET) derived from country thresholds (table S1) (*26*). Five sets of predictive variables, namely climatic, geological, soil textural, topographic, and socioeconomic, were included as proxies of natural and anthropogenic processes governing metal abundance in soil. Extremely randomized trees (ERT) was selected as the best-performing machine learning model (*27*). The models were validated using an independent data set, which verified high model precision and accuracy unrelated to numerical overfitting. The models were then used to project data onto a soil pollution map on a global scale, excluding any permafrost and desert areas (*26*).

Globally, our model estimates that 14% to 17% (95% confidence interval) of surface soils exceed the AT for at least one toxic metal in cropland areas (Fig. 1). Probabilities of individual metal exceedance vary geographically (fig S4-S10). The global exceedance rate of Cd is the highest, reaching 9.0% (-1.9%/+1.5%). Cadmium exceedance for agricultural soil is the most notable in northern and central India, Pakistan, Bangladesh, southern China, southern parts of Thailand and Cambodia, Iran, Turkey, Ethiopia, Nigeria, South Africa, Mexico, and Cuba. Both anthropogenic sources and geogenic enrichment likely contributed to the elevated Cd concentrations in these regions (*6*, *8*, *28*, *29*). The exceedance rates of Ni and Cr reach 5.8% (-1.8%/+1.1%) and 3.2% (-0.7%/+1.6%), respectively. Their exceedance is the most prevalent in Middle-East, subarctic Russia, and eastern Africa, likely due to high geogenic background as well as mining activities (*28*, *30*). Soil As exceedance occurred at a rate of 1.1% (-0.04%/+0.3%), and was the most notable in southern and southwestern China, south and Southeast Asia, West Africa, and central parts of South America, which coincide with observed and predicted areas of high As concentration in groundwater (*14*). The exceedance rate of Co is 1.1% (-0.1%/+2.9%), and was the most prevalent in Zambia, the Democratic Republic of the Congo, and Ethiopia, likely due to mining related activities (*31*). Globally 6.8% (-1.7%/+1.9%) of surficial soil exceeded HHET, with a similar or smaller exceedance than AT exceedance owning to generally less stringent threshold values (Fig 2, figs S11-S17).

Soil pollution by toxic metals has significant impacts on food production and food safety. We estimate that 242 million ha (-26/+27 million ha), or 16% of global cropland is affected by toxic metal exceedances. Among the areas most at risk, southern China, northern and central India, and the mid-East, are well documented to have elevated toxic metal concentrations in their soils (*32*-*34*). Limited data exist for Africa and the prediction will require more soil sampling and analysis for verification (*35*).

By overlaying the human health and ecological risk map over global population distribution in 2020, it is estimated that 0.9-1.4 billion people live in the high-risk areas (Fig 2B). However, it should be noted that the actual risks posed by soil metals are dependent upon their mobility, overall bioavailability, and human exposure pathway dynamics (*36*, *37*). Exposure and toxic effects also depend on individual dietary habits and food deprivation, as well as the degree of co-occurrence of multiple elements (Fig. 2C). Moreover, international trade of food products originating from high risk countries may lead to a spill-over effect and dispersion of such risks (Fig. 1D).

Our study identified a notable high-risk zone in low-latitude Eurasia and across southern Europe, the mid-East, South Asia, and southern China. This belt coincides with the geographical distribution of several ancient cultures, including ancient Greek civilizations, the Roman Empire, Persian culture, ancient India, and Yangtze-river Chinese culture (fig. S25). This intercontinental “metal-enriched corridor” is attributed to a combination of anthropogenic and environmental factors (discussed below). Since metals do not degrade, this zone can be regarded as a keystone indicator of the Anthropocene era.

Natural and Anthropogenic Drivers

Several environmental drivers affect the global distribution of toxic metal exceedances. Near-surface temperature, precipitation, and potential evapotranspiration have the strongest positive effects (*38*), likely contributing to relatively high metal exceedance in southern China, India, mid-East, Central America, and Central Africa. Such conditions accelerate the weathering processes that release metals from soil parent materials and enhance the enrichment of metals in clay minerals and iron- or aluminum-oxides (*22*). In contrast, the frequency of ground frosts and wet day frequencies show the strongest negative effects (*38*). This may be due to weak weathering-induced influx and strong leaching-related efflux of metals (*39*), as well as weak plant-pump effects limiting vertical enrichment (*17*). The subtropical monsoon climate zones, which are important for global agriculture, tend to be hot and humid despite the dry season. This climate zone has a metal exceedance rate of 34% (-5%/+4%) for the AT, significantly higher than the global average of 16% (-2%/+2%). In contrast, the metal exceedance rate in the cold and humid hemi-boreal climate zone is much lower at 6.0% (-2.4%/+5.5%) (Fig. 3B). We also found that high elevation and steep slope landscapes correspond to more prevalent metal exceedance (Fig. 3E-G) owing to the topography affecting rock weathering, soil formation and erosion, and therefore influencing the leaching and accumulation of metals (*40-42*). In mountainous areas with a low percentage of flat areas and high percentage of steep slopes, the metal exceedance rate is 15% (-4%/+2%) for HHET and 29% (-1%/+3%) for AT, nearly twice the global averages.

Socioeconomic factors are also important drivers governing global toxic metal distribution patterns. Proxies of mining intensity, as identified by ore/metal exports, mineral rents, mineral depletion, and ores/metal imports, were the strongest socioeconomic predictors of toxic metal exceedances, highlighting the major contribution of mining and smelting on metal accumulation in soils at a global scale (*6*, *43*, *44*). The proportion of irrigated land was also found to be a strong predictor of metal exceedance, consistent with previous reports that irrigation water contaminated by industrial activities can cause widespread contamination of agricultural soils (*6*, *8*, *19*). In areas with intensive mining activities and a high percentage of surface irrigation (Fig. 3I-L), the metal exceedance rate was 17% (-5%/+4%) for HHET and 36% (-7%/+4%) for AT, more than twice the global average. Although irrigation with groundwater extracted from arsenic-bearing aquifers in the region south of Himalayas resulted in hot spots of As in soils (*8*), in general, the use of groundwater for irrigation is a strong predictor of toxic metal non-exceedance on a global scale. This suggests that groundwater may generally contain lower levels of toxic metals than other irrigation water sources, thus serving as a carrier of metal efflux rather than influx, except in areas with high geogenic background or serious anthropogenic pollution (*45*).

We used structural equation models (SEM) to assess the causal links between irrigation, mining, plant pumping, weathering, leaching, and exceedance rate and hazard level (Fig. 4A, B, fig. S22) and found that weathering and plant pumping contribute substantially to the concentrations of As, Cd, Co, Cu in soil. Furthermore, SEM results verified that anthropogenic processes including mining and irrigation provided significant contributions for most of the toxic metals. Although many effects are exerted via direct influencing pathways, a significant portion of the influences may be exerted indirectly (Fig. 4C). Indirect pathways account for 96%, 87%, 62%, and 62% of the net effect of mining on hazard level for As, Cd, Co, and Cu. These SEM results were in good accordance with the complex importance features of the machine learning models (Fig. 4D), and support our hypothesis that soil toxic metal enrichment is governed by the interplay of a wide range of biogeophysical and socioeconomic variables at broad spatiotemporal scales.

Discussion

Our model results show that soil contamination is occurring on a global scale, posing significant risks to both ecosystems and human health (*7, 46*), and threatening water quality and food security (*6, 8*). The model prediction includes both known soil pollution areas and previously undocumented areas of concern (fig. S23-S24). Some of these regions, such as Southern China and the Middle East, have been reported previously, but we were able to delineate the risk zones continuously on a global scale. Our machine learning models used data from the public domain to provide an assessment of regional soil pollution, and the results show that the technique is a useful screening tool that can complement traditional soil pollution mapping methods. There is an ongoing global initiatives on soil pollution prevention and restoration under the United Nations Environment Programme (UNEP) and the FAO (*35, 47*). Our results suggest that international aid should be allocated to facilitate soil pollution surveys in data-sparse regions such as Sub-Saharan Africa.

Recent large-scale studies in Europe found a mysterious trend North of the 55^{o} latitude line, which demarcates high-metal soils in the south from the low-metal soils in the north (*11, 48*). This phenomenon had been attributed to the coincidental match with the maximum extent of the last glaciation; however, the overall mechanism and drivers remain unclear. Our results now reveal that the toxic metal-enriched area across southern Europe is part of a more extensive trans-continental metal-enriched corridor spanning across low latitude Eurasia (Fig. 1A). We postulate that this corridor of long-lasting legacy of human influence was formed due to strong weathering of metal-enriched parent rocks (*12, 49*) and plant-pumping effects (*13, 17*), a lower degree of leaching associated with precipitation and terrain (*12*), and a long-history of mining and smelting activities occurring since ancient civilizations began (*8*).

Our models were validated using a series of uncertainty analyses (*26*) (figs. S18-S21). Mapping the extent of spatial extrapolation showed that our dataset provides a good coverage of most environmental conditions, with the least represented pixels and highest proportion of extrapolation in Southeast Asia, Russia, and Africa. Due to lack of sampling data in developing countries and remote regions, our model still has relatively high degrees of uncertainty in northern Russia, central India, and Africa (fig. S2). Moreover, metal concentrations in soil have high spatial heterogeneity and may vary significantly over short distances. The present study is based on average metal concentrations on a 10-km grid, which is more reflective of diffusive and regional pollution rather than site specific conditions. The data may be sufficient for risk screening purposes but are inadequate to support risk mitigation. Soil remediation needs to rely upon site-specific delineation of lateral and vertical extent of soil pollution, as well as a better understanding of metal sources, fate and transport dynamics, and bioavailability (*12*).

Soil pollution can have a profound impact on global food security and public health. For the millions of people making a living on the 14% to 17% of globally polluted cropland, the bioaccumulation of toxic metals in crops and farm animals can affect biodiversity and productivity, cause detrimental health effects, and exacerbate poverty. The collateral effects on the global food chain are unknown at this time, especially in the context of how global trade dynamics may affect the distribution of contaminated agricultural products. These large areas of toxic metal enrichment are expected to continue to increase due to the growth in demand for critical metals required to support the net zero 'green transition' and the development of photovoltaic devices, wind turbines, and electric vehicle batteries (*50, 51*). We hope that the global soil pollution data presented in this report will serve as scientific alert for policy makers and farmers to take immediate and necessary measures to better protect the world's precious soil resources.

References and Notes

1. FAO, Healthy soils are the basis for healthy food production (Food and Agriculture Organization of the United Nations, Rome, Italy, 2015).
2. M. van Dijk, T. Morley, M. L. Rau, Y. Saghai, A meta-analysis of projected global food demand and population at risk of hunger for the period 2010–2050. *Nat. Food* **2**, 494-501 (2021).
3. UNCCD, Global land outlook (United Nations Convention to Combat Desertification, Bonn, Germany, 2017).
4. FAO, Status of the world's soil resources (Food and Agriculture Organization of the United Nations, Rome, Italy, 2015).
5. FAO, Saving our soils by all earthly ways possible (Food and Agriculture Organization of the United Nations, Rome, Italy, 2022).
6. D. Hou, D. O'Connor, A. D. Igalavithana, D. S. Alessi, J. Luo, D. C. W. Tsang, D. L. Sparks, Y. Yamauchi, J. Rinklebe, Y. S. Ok, Metal contamination and bioremediation of agricultural soils for food safety and sustainability. *Nat. Rev. Earth Environ.* **1**, 366–381 (2020).
7. O. Coban, G. B. De Deyn, M. van der Ploeg, Soil microbiota as game-changers in restoration of degraded lands. *Science* **375**, eabe0725 (2022).
8. FAO and UNEP, Global assessment of soil pollution (Food and Agriculture Organization of the United Nations and United Nations Environment Programme, Rome, Italy, 2021).
9. A. Kabata-Pendias, *Trace Elements in Soils and Plants* (CRC Press, 2000).
10. MEP, National Soil Contamination Survey Report (Ministry of Environmental Protection of the People's Republic of China, Beijing, China, 2014).
11. G. Tóth, T. Hermann, G. Szatmári, L. Pásztor, Maps of heavy metals in the soils of the European Union and proposed priority areas for detailed assessment. *Sci. Total Environ.* **565**, 1054-1062 (2016).
12. B. J. Alloway, Ed., *Heavy Metals in Soils: Trace Metals and Metalloids in Soils and their Bioavailability* (Springer Dordrecht, ed. 3, 2013).
13. A. A. Meharg, C. Meharg, The pedosphere as a sink, source, and record of anthropogenic and natural arsenic atmospheric deposition. *Environ. Sci. Technol.* **55**, 7757-7769 (2021).
14. J. Podgorski, M. Berg, Global threat of arsenic in groundwater. *Science* **368**, 845-850 (2020).
15. J. O. Nriagu, A global assessment of natural sources of atmospheric trace metals. *Nature* **338**, 47-49 (1989).

16. X. Liu, C. Tournassat, S. Grangeon, A. G. Kalinichev, Y. Takahashi, M. Marques Fernandes, Molecular-level understanding of metal ion retention in clay-rich materials. *Nat. Rev. Earth Environ.* **3**, 461-476 (2022).
17. M. Imseng, M. Wiggenhauser, A. Keller, M. Müller, M. Rehkämper, K. Murphy, K. Kreissig, E. Frossard, W. Wilcke, M. Bigalke, Fate of Cd in agricultural soils: a stable isotope approach to anthropogenic impact, soil formation, and soil-plant cycling. *Environ. Sci. Technol.* **52**, 1919-1928 (2018).
18. S. Hong, J.-P. Candelone, C. C. Patterson, C. F. Boutron, History of ancient copper smelting pollution during Roman and medieval times recorded in Greenland ice. *Science* **272**, 246-249 (1996).
19. M. G. Macklin, C. J. Thomas, A. Mudbhatkal, P. A. Brewer, K. A. Hudson-Edwards, J. Lewin, P. Scussolini, D. Eilander, A. Lechner, J. Owen, G. Bird, D. Kemp, K. R. Mangalaa, Impacts of metal mining on river systems: a global assessment. *Science* **381**, 1345-1350 (2023).
20. S. Hong, J.-P. Candelone, C. C. Patterson, C. F. Boutron, Greenland ice evidence of hemispheric lead pollution two millennia ago by Greek and Roman civilizations. *Science* **265**, 1841-1843 (1994).
21. J. O. Nriagu, J. M. Pacyna, Quantitative assessment of worldwide contamination of air, water and soils by trace metals. *Nature* **333**, 134 (1988).
22. M. Imseng, M. Wiggenhauser, M. Müller, A. Keller, E. Frossard, W. Wilcke, M. Bigalke, The fate of Zn in agricultural soils: a stable isotope approach to anthropogenic impact, soil formation, and soil–plant cycling. *Environ. Sci. Technol.* **53**, 4140-4149 (2019).
23. L. R. Lado, T. Hengl, H. I. Reuter, Heavy metals in European soils: a geostatistical analysis of the FOREGS Geochemical database. *Geoderma* **148**, 189-199 (2008).
24. S. Maas, R. Scheifler, M. Benslama, N. Crini, E. Lucot, Z. Brahmia, S. Benyacoub, P. Giraudoux, Spatial distribution of heavy metal concentrations in urban, suburban and agricultural soils in a Mediterranean city of Algeria. *Environ. Pollut.* **158**, 2294-2301 (2010).
25. Y. Hu, H. Cheng, Application of Stochastic Models in Identification and Apportionment of Heavy Metal Pollution Sources in the Surface Soils of a Large-Scale Region. *Environ. Sci. Technol.* **47**, 3752-3760 (2013).
26. Materials and methods are available as supplementary materials.
27. P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees. *Mach. Learn.* **63**, 3-42 (2006).
28. BGS, World Mineral Production 2016-2020 (British Geological Survey, Nottingham, UK, 2022).
29. A. Kubier, R. T. Wilkin, T. Pichler, Cadmium in soils and groundwater: a review. *Appl. Geochem.* **108**, 104388 (2019).
30. BGS, Mineral profile: nickel (British Geological Survey, Nottingham, UK, 2008).
31. G. Gunn, Ed., *Critical Metals Handbook* (John Wiley & Sons, 2013).
32. H. Chen, Y. Teng, S. Lu, Y. Wang, J. Wang, Contamination features and health risk of soil heavy metals in China. *Sci. Total Environ.* **512**, 143-153 (2015).
33. N. Gupta, K. K. Yadav, V. Kumar, S. Krishnan, S. Kumar, Z. D. Nejad, M. M. Khan, J. Alam, Evaluating heavy metals contamination in soil and vegetables in the region of North India: Levels, transfer and potential human health risk analysis. *Environ. Toxicol. Pharmacol.* **82**, 103563 (2021).

34. M. Amini, M. Afyuni, N. Fathianpour, H. Khademi, H. Flühler, Continuous soil pollution mapping using fuzzy logic and spatial interpolation. *Geoderma* **124**, 223-233 (2005).
35. UNEA, Managing soil pollution to achieve Sustainable Development (United Nations Environment Assembly, Nairobi, Kenya, 2018).
36. G. Liu, J. Wang, X. Liu, X. Liu, X. Li, Y. Ren, J. Wang, L. Dong, Partitioning and geochemical fractions of heavy metals from geogenic and anthropogenic sources in various soil particle size fractions. *Geoderma* **312**, 104-113 (2018).
37. G. Lin, C. Zhang, Z. Yang, Y. Li, C. Liu, L. Q. Ma, High geological background concentrations of As and Cd in karstic soils may not contribute to greater risks to human health via rice consumption. *J. Hazard. Mater.* **480**, 135876 (2024).
38. D. Hou, X. Jia, L. Wang, S. P. McGrath, Y.-G. Zhu, Q. Hu, F.-J. Zhao, M. S. Bank, D. O'Connor, J. Nriagu, Global soil pollution by toxic metals threatens agriculture and human health [Dataset]. Dryad (2025). https://doi.org/10.5061/dryad.83bk3jb2z.
39. F. Meite, P. Alvarez-Zaldívar, A. Crochet, C. Wiegert, S. Payraudeau, G. Imfeld, Impact of rainfall patterns and frequency on the export of pesticides and heavy-metals from agricultural soils. *Sci. Total Environ.* **616-617**, 500-509 (2018).
40. W. E. Dietrich, J. T. Perron, The search for a topographic signature of life. *Nature* **439**, 411-418 (2006).
41. C. Boente, D. Baragaño, N. García-González, R. Forján, A. Colina, J. R. Gallego, A holistic methodology to study geochemical and geomorphological control of the distribution of potentially toxic elements in soil. *Catena* **208**, 105730 (2022).
42. Q. Ding, G. Cheng, Y. Wang, D. Zhuang, Effects of natural factors on the spatial distribution of heavy metals in soils surrounding mining regions. *Sci. Total Environ.* **578**, 577-585 (2017).
43. A. S. Vega, G. Arce, J. I. Rivera, S. E. Acevedo, S. Reyes-Paecke, C. A. Bonilla, P. Pastén, A comparative study of soil metal concentrations in Chilean urban parks using four pollution indexes. *Appl. Geochem.* **141**, 105230 (2022).
44. E. Saljnikov, V. Mrvić, D. Čakmak, D. Jaramaz, V. Perović, S. Antić-Mladenović, P. Pavlović, Pollution indices and sources appointment of heavy metal pollution of agricultural soils near the thermal power plant. *Environ. Geochem. Health* **41**, 2265-2279 (2019).
45. T. Gleeson, M. Cuthbert, G. Ferguson, D. Perrone, Global groundwater sustainability, resources, and systems in the Anthropocene. *Annu. Rev. Earth Planet. Sci.* **48**, 431-463 (2020).
46. UNEP, Towards a Pollution-Free Planet (United Nations Environment Programme, Nairobi, Kenya, 2017).
47. FAO, Soil pollution: a hidden reality (Food and Agriculture Organization of the United Nations, Rome, Italy, 2018).
48. C. Reimann, P. de Caritat, G. P. Team, N. P. Team, New soil composition data for Europe and Australia: demonstrating comparability, identifying continental-scale processes and learning lessons for global geochemical mapping. *Sci. Total Environ.* **416**, 239-252 (2012).
49. H. S. Moghadam, W. L. Griffin, X. H. Li, J. F. Santos, O. Karsli, R. J. Stern, G. Ghorbani, S. Gain, R. Murphy, S. Y. O'Reilly, Crustal evolution of NW Iran: Cadomian Arcs, Archean fragments and the Cenozoic magmatic flare-up. *J. Petrol.* **58**, 2143-2190 (2017).

50. K. Bhuwalka, F. R. Field III, R. D. De Kleine, H. C. Kim, T. J. Wallington, R. E. Kirchain, Characterizing the Changes in Material Use due to Vehicle Electrification. *Environ. Sci. Technol.* **55**, 10097-10107 (2021).
51. A. Farina, A. Anctil, Material consumption and environmental impact of wind turbines in the USA and globally. *Resour. Conserv. Recycl.* **176**, 105938 (2022).
52. K. Jomova, Z. Jenisova, M. Feszterova, S. Baros, J. Liska, D. Hudecova, C. Rhodes, M. Valko, Arsenic: toxicity, oxidative stress and human disease. *J. Appl. Toxicol.* **31**, 95-107 (2011).
53. WHO, Arsenic primer: Guidance on the investigation and mitigation of arsenic contamination (World Health Organization, Geneva, Switzerland, 2018).
54. ATSDR, ATSDR’s Substance Priority List (Agency for Toxic Substances and Disease Registry, Atlanta, Georgia, 2020).
55. D. Wang, P. Luo, Z. Zou, Q. Wang, M. Yao, C. Yu, S. Wei, B. Sun, K. Zhu, Q. Zeng, J. Li, B. Liang, A. Zhang, Alterations of arsenic levels in arsenicosis residents and awareness of its risk factors: A population-based 20-year follow-up study in a unique coal-borne arsenicosis County in Guizhou, China. *Environ. Int.* **129**, 18-27 (2019).
56. L. M. Camacho, M. Gutiérrez, M. T. Alarcón-Herrera, M. D. L. Villalba, S. Deng, Occurrence and treatment of arsenic in groundwater and soil in northern Mexico and southwestern USA. *Chemosphere* **83**, 211-225 (2011).
57. C. Ferreccio, A. M. Sancha, Arsenic exposure and its impact on health in Chile. *J. Health Popul. Nutr.* **24**, 164-175 (2006).
58. WHO, Exposure to Cadmium: A Major Public Health Concern (World Health Organization, Geneva, Switzerland, 2019).
59. Y. Hu, H. Cheng, S. Tao, The challenges and solutions for cadmium-contaminated rice in China: a critical review. *Environ. Int.* **92**, 515-532 (2016).
60. A. Luch, Ed., *Molecular, Clinical and Environmental Toxicology Volume 3: Environmental Toxicology* (Springer, 2012).
61. L. Leyssens, B. Vinck, C. Van Der Straeten, F. Wuyts, L. Maes, Cobalt toxicity in humans—A review of the potential sources and systemic health effects. *Toxicology* **387**, 43-56 (2017).
62. L. M. Gaetke, C. K. Chow, Copper toxicity, oxidative stress, and antioxidant nutrients. *Toxicology* **189**, 147-163 (2003).
63. B. Shahzad, M. Tanveer, A. Rehman, S. A. Cheema, S. Fahad, S. Rehman, A. Sharma, Nickel; whether toxic or essential for plants and environment-A review. *Plant Physiol. Biochem.* **132**, 641-651 (2018).
64. ATSDR, Toxicological Profile for Nickel (Agency for Toxic Substances and Disease Registry, Atlanta, Georgia, 2005).
65. WHO, Lead poisoning and health (World Health Organization, Geneva, Switzerland, 2017).
66. B. P. Lanphear, R. Hornung, J. Khoury, K. Yolton, P. Baghurstl, D. C. Bellinger, R. L. Canfield, K. N. Dietrich, R. Bornschein, T. Greene, S. J. Rothenberg, H. L. Needleman, L. Schnaas, G. Wasserman, J. Graziano, R. Roberts, Low-level environmental lead exposure and children's intellectual function: An international pooled analysis. *Environ. Health Perspect.* **113**, 894-899 (2005).
67. B. P. Lanphear, S. Rauch, P. Auinger, R. W. Allen, R. W. Hornung, Low-level lead exposure and mortality in US adults: a population-based cohort study. *Lancet Public Health* **3**, e177-e184 (2018).

68. D. Hou, D. O'Connor, P. Nathanail, L. Tian, Y. Ma, Integrated GIS and multivariate statistical analysis for regional scale assessment of heavy metal soil contamination: A critical review. *Environ. Pollut.* **231**, 1188-1200 (2017).
69. J. Balesdent, I. Basile-Doelsch, J. Chadoeuf, S. Cornu, D. Derrien, Z. Fekiacova, C. Hatte, Atmosphere-soil carbon transfer as a function of soil depth. *Nature* **559**, 599-602 (2018).
70. V. Iñigo, M. Andrades, J. Alonso-Martirena, A. Marín, R. Jiménez-Ballesta, Multivariate statistical and GIS-based approach for the identification of Mn and Ni concentrations and spatial variability in soils of a humid Mediterranean environment: La Rioja, Spain. *Water Air Soil Pollut.* **222**, 271-284 (2011).
71. J. Hartmann, N. Moosdorf, Global Lithological Map Database v1.0 (gridded to 0.5° spatial resolution) [Dataset]. PANGAEA (2012). https://doi.org/10.1594/PANGAEA.788537.
72. H. T. Davis, C. M. Aelion, S. McDermott, A. B. Lawson, Identifying natural and anthropogenic sources of metals in urban and rural soils using GIS-based data, PCA, and spatial interpolation. *Environ. Pollut.* **157**, 2378-2385 (2009).
73. I. Harris, T. J. Osborn, P. Jones, D. Lister, Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset. *Sci. Data* **7**, 109 (2020).
74. Y. Q. Zhang, J. L. Pena-Arancibia, T. R. McVicar, F. H. S. Chiew, J. Vaze, C. M. Liu, X. J. Lu, H. X. Zheng, Y. P. Wang, Y. Y. Liu, D. G. Miralles, M. Pan, Multi-decadal trends in global terrestrial evapotranspiration and its components. *Sci. Rep.* **6**, 12 (2016).
75. M. H. Ali, A.-R. A. Mustafa, A. A. El-Sheikh, Geochemistry and spatial distribution of selected heavy metals in surface soil of Sohag, Egypt: a multivariate statistical and GIS approach. *Environ. Earth Sci.* **75**, 1257 (2016).
76. A. Facchinelli, E. Sacchi, L. Mallen, Multivariate statistical and GIS-based approach to identify heavy metal sources in soils. *Environ. Pollut.* **114**, 313-324 (2001).
77. X. Liu, J. Wu, J. Xu, Characterizing the risk assessment of heavy metals and sampling uncertainty analysis in paddy field by geostatistics and GIS. *Environ. Pollut.* **141**, 257-264 (2006).
78. A. Mihailović, L. Budinski-Petković, S. Popov, J. Ninkov, J. Vasin, N. Ralević, M. V. Vasić, Spatial distribution of metals in urban soil of Novi Sad, Serbia: GIS based approach. *J. Geochem. Explor.* **150**, 104-114 (2015).
79. FAO, Harmonized World Soil Database Version 1.2 (Food and Agriculture Organization of the United Nations, Rome, Italy, 2012).
80. M. E. Kylander, A. M. Cortizas, S. Rauch, D. J. Weiss, Lead penetration and leaching in a complex temperate soil profile. *Environ. Sci. Technol.* **42**, 3177-3184 (2008).
81. R. B. Kheir, M. H. Greve, C. Abdallah, T. Dalgaard, Spatial soil zinc content distribution from terrain parameters: A GIS-based decision-tree model in Lebanon. *Environ. Pollut.* **158**, 520-528 (2010).
82. G. Fischer, F. Nachtergaele, S. Prieler, H. T. van Velthuizen, L. Verelst, D. Wiberg, Global Agro-ecological Zones Assessment for Agriculture (GAEZ 2008) (IIASA, Laxenburg, Austria and FAO, Rome, Italy, 2008).
83. S. R. Gaffin, X. Xing, G. Yetman, Global 15 x 15 Minute Grids of the Downscaled GDP Based on the SRES B2 Scenario, 1990 and 2025 (version 1.00) [Dataset]. NASA Socioeconomic Data and Applications Center (2004). https://doi.org/10.7927/H4NC5Z4X.

84. Center For International Earth Science Information Network-CIESIN-Columbia University, Gridded Population of the World, version 4 (GPWv4): Population Density Adjusted to Match 2015 Revision UN WPP Country Totals, Revision 11 (version 4.11) [Dataset]. NASA Socioeconomic Data and Applications Center (2018). https://doi.org/10.7927/H4F47M65.
85. Y. Sun, Q. Zhou, X. Xie, R. Liu, Spatial, sources and risk assessment of heavy metal contamination of urban soils in typical regions of Shenyang, China. *J. Hazard. Mater.* **174**, 455-462 (2010).
86. Y. Shan, M. Tysklind, F. Hao, W. Ouyang, S. Chen, C. Lin, Identification of sources of heavy metals in agricultural soils using multivariate analysis and GIS. *J. Soils Sediments* **13**, 720-729 (2013).
87. J. A. R. Martín, M. L. Arias, J. M. G. Corbí, Heavy metals contents in agricultural topsoils in the Ebro basin (Spain). Application of the multivariate geoestatistical methods to study spatial variations. *Environ. Pollut.* **144**, 1001-1012 (2006).
88. P. Potter, N. Ramankutty, E. M. Bennett, S. D. Donner, Global Fertilizer and Manure, version 1: Phosphorus in Manure Production (version 1.00) [Dataset]. NASA Socioeconomic Data and Applications Center (2011). https://doi.org/10.7927/H49Z92TD.
89. S. Siebert, V. Henrich, K. Frenken, J. Burke, Global Map of Irrigation Areas (version 5) [Dataset]. Rheinische Friedrich-Wilhelms-University and the Food and Agriculture Organization of the United Nations (2013). https://www.fao.org/aquastat/en/geospatial-information/global-maps-irrigated-areas/latest-version/.
90. I. Manisalidis, E. Stavropoulou, A. Stavropoulos, E. Bezirtzoglou, Environmental and Health Impacts of Air Pollution: A Review. *Front. Public Health* **8**, 14 (2020).
91. F. A. Nicholson, S. R. Smith, B. Alloway, C. Carlton-Smith, B. Chambers, An inventory of heavy metals inputs to agricultural soils in England and Wales. *Sci. Total Environ.* **311**, 205-219 (2003).
92. D. Tong, Q. Zhang, S. J. Davis, F. Liu, B. Zheng, G. Geng, T. Xue, M. Li, C. Hong, Z. Lu, D. G. Streets, D. Guan, K. He, Targeted emission reductions from global super-polluting power plant units. *Nat. Sustain.* **1**, 59-68 (2018).
93. C. Carlon, Derivation methods of soil screening values in Europe. A review and evaluation of national procedures towards harmonisation (European Commission, Joint Research Centre, Ispra, 2007).
94. CCME, Canadian Environmental Quality Guidelines (Canadian Council of Ministers of the Environment, Winnipeg, Canada, 2015).
95. MEE, Soil Environmental Quality: Risk Control Standard for Soil Contamination of Agricultural Land (Ministry of Ecology and Environment of the People's Republic of China GB 15618-2018, Beijing, China, 2018).
96. MEE, Soil Environmental Quality: Risk Control Standard for Soil Contamination of Development Land (Ministry of Ecology and Environment of the People's Republic of China GB 36600-2018, Beijing, China, 2018).
97. USEPA, Regional Screening Table User's Guide (United States Environmental Protection Agency, Washington, D. C., 2020).
98. C. Walck, Hand-book on statistical distributions for experimentalists (Stockholm University, Stockholm, Sweden, 1996).
99. L. Rodríguez-Lado, G. Sun, M. Berg, Q. Zhang, H. Xue, Q. Zheng, C. A. Johnson, Groundwater arsenic contamination throughout China. *Science* **341**, 866-868 (2013).

100. G. Tóth, T. Hermann, M. Da Silva, L. Montanarella, Heavy metals in agricultural soils of the European Union with implications for food safety. *Environ. Int.* **88**, 299-309 (2016).
101. N. S.-N. Lam, Spatial interpolation methods: a review. *The American Cartographer* **10**, 129-150 (1983).
102. A. M. Ellison, Bayesian inference in ecology. *Ecol. Lett.* **7**, 509-520 (2004).
103. W. A. Link, R. J. Barker, *Bayesian Inference: With Ecological Applications* (Academic Press, 2009).
104. W. R. Tobler, A computer movie simulating urban growth in the Detroit region. *Econ. Geogr.* **46**, 234-240 (1970).
105. G. E. Box, G. C. Tiao, *Bayesian Inference in Statistical Analysis* (John Wiley & Sons, 2011).
106. D. I. MacKenzie, J. D. Nichols, J. A. Royle, K. H. Pollock, L. Bailey, J. E. Hines, *Occupancy Estimation and Modeling: Inferring Patterns and Dynamics of Species Occurrence* (Elsevier, 2017).
107. L. Breiman, Random forests. *Mach. Learn.* **45**, 5-32 (2001).
108. G. Biau, E. Scornet, A Random Forest Guided Tour. *Test* **25**, 197-227 (2015).
109. C. Desir, C. Petitjean, L. Heutte, M. Salaun, L. Thiberville, Classification of endomicroscopic images of the lung based on random subwindows and extra-trees. *IEEE Trans. Biomed. Eng.* **59**, 2677 (2012).
110. G. Chandrashekar, F. Sahin, A survey on feature selection methods. *Comput. Electr. Eng.* **40**, 16-28 (2014).
111. B. F. Darst, K. C. Malecki, C. D. Engelman, Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genet.* **19**, 65 (2018).
112. S. McGee, *Evidence-Based Physical Diagnosis* (Elsevier, ed. 4, 2018).
113. F. E. Nelson, O. A. Anisimov, N. I. Shiklomanov, Current Permafrost Distribution in the Northern Hemisphere [Dataset]. National Science Foundation Arctic Systems Science Program (2000). https://databasin.org/datasets/6893ca9aaee042ea83899ada60219665/.
114. T. Patterson, N. V. Kelso, World Land-Based Polygon Features, 1:10 million [Dataset]. North American Cartographic Information Society (2012). https://purl.stanford.edu/bh326sc0899.
115. S. Lundberg, S. I. Lee, A Unified Approach to Interpreting Model Predictions (Neural Information Processing Systems, 2017).
116. A. S. Antonini, J. Tanzola, L. Asiain, G. R. Ferracutti, S. M. Castro, E. A. Bjerg, M. L. Ganuza, Machine Learning model interpretability using SHAP values: Application to Igneous Rock Classification task. *Appl. Comput. Geosci.* **23**, 100178 (2024).
117. E. A. Freeman, G. G. Moisen, A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecol. Model.* **217**, 48-58 (2008).
118. J. Gao, Downscaling global spatial population projections from 1/8-degree to 1-km grid cells (National Center for Atmospheric Research, Boulder, USA, 2017).
119. P. M. Shankar, Tutorial overview of simple, stratified, and parametric bootstrapping. *Eng. Rep.* **2**, e12096 (2020).
120. L. Six, E. Smolders, Future trends in soil cadmium concentration under current cadmium fluxes to European agricultural soils. *Sci. Total Environ.* **485**, 319-328 (2014).
121. J. B. Grace, *Structural equation modeling and natural systems* (Cambridge Univ. Press, 2006).

122. USEPA, Risk assessment guidance for Superfund. Volume I: human health evaluation manual (Part A) (United States Environmental Protection Agency, Washington, D. C., 1989).
123. M. Delgado-Baquerizo, F. T. Maestre, A. Gallardo, M. A. Bowker, M. D. Wallenstein, J. L. Quero, V. Ochoa, B. Gozalo, M. García-Gómez, S. Soliveres, Decoupling of soil nutrient cycles as a function of aridity in global drylands. *Nature* **502**, 672-676 (2013).
124. D. A. Kenny, B. Kaniskan, D. B. McCoach, The performance of RMSEA in models with small degrees of freedom. *Sociol. Method. Res.* **44**, 486-507 (2015).
125. M. Oliver, R. Webster, A tutorial guide to geostatistics: Computing and modelling variograms and kriging. *Catena* **113**, 56-69 (2014).
126. USEPA, Supplemental guidance for developing soil screening levels for superfund sites (United States Environmental Protection Agency, Washington, D. C., 2002).
127. USEPA, Exposure factors handbook (United States Environmental Protection Agency, Washington, D. C., 2011).
128. USDOE, The Risk Assessment Information System (RAIS) (United States Department of Energy, Washington, D. C., 2011).

Acknowledgments:

We thank the anonymous reviewers for their constructive comments, and the many providers of data used in our models. **Funding:** National Natural Science Foundation of China grant 42225703 (DH); National Key Research and Development Program of China grant 2020YFC1808000 (DH).

Author contributions:

Conceptualization: DH, XJ, SPM, YZ, QH, FZ, MSB, DO, JN

Methodology: DH, XJ, LW, SPM, FZ, DO

Investigation: DH, XJ, LW

Visualization: DH, XJ

Funding acquisition: DH

Project administration: DH, LW

Supervision: DH

Writing – original draft: DH, XJ

Writing – review & editing: DH, SPM, YZ, QH, FZ, MSB, DO, JN

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: Data and code generated during this study is publicly available and can be accessed at (*38*).

Supplementary Materials

Materials and Methods

Figs. S1 to S27

Tables S1 to S10

References (*52–128*)

Fig. 1. Global soil pollution by toxic metals exceeding agricultural threshold (AT). (**A**) Aggregate distribution of exceedance of arsenic, cadmium, cobalt, chromium, copper, nickel, and lead; color code shows the maximum probability of exceedance among the seven metals. (**B-C**) zoomed-in sections of globally important food production areas. (D) Predicted Cd exceedance rates and average soil pH indicative of Cd mobility in the major rice export countries. Country abbreviation: IN = India, TH = Thailand, VN = Vietnam, PK = Pakistan, CN = China, US = United States, BR = Brazil, PY = Paraguay, EU = European Union, AR = Argentina.

Fig. 2. Global distribution of soil toxic metals exceeding human health and ecological threshold (HHET). (**A**) Map of metal concentration exceedance. (**B**) Population density in areas with >0.5 probability of metal exceeding ecological and human health threshold. (**C**) Combined soil pollution by toxic metals, with line width in the Sankey diagram showing the proportion of all dual comingled pollution. (**D**) Density histogram showing the relative frequency of exceedance probability of various continents, adjusted by area of each continent.

Fig. 3. Natural and anthropogenic drivers of soil metal exceedance. (**A**) Global distribution of subtropical monsoon (SM) and hemiboreal (HB) climate zones. (**B**) Exceedance rate in global, SM, and HB climate zones. (**C**) Exceedance rate increases as precipitation increases. (**D**) Exceedance rate decreases as ground frost frequency increases. (**E**) Global distribution of hilly mountain areas (HMA), with <2% of area sloped between 0.005 and 0.02, and >10% of area sloped between 0.3 and 0.45, and elevation >1,000 meter above mean sea level. (**F**) Exceedance rate in HMA is significantly higher than global average. (**G**) Exceedance rate decreases as proportion of flat land increases. (**H**) Exceedance rate increases as elevation increases. (**I**) Global distribution of irrigated and mineral rich regions (IMR), with proportion of irrigation exceeding 90% and ores and metals imports over 5% of merchandise imports (MI). (**J**) Exceedance rate in IMR compared with global average. (**K**) Exceedance rate increases as the proportion of irrigation increases. (**L**) Exceedance rate increases as the proportion of ores and metals imports increases. Regression lines are shown in C, D, G, H, K, L, with "L" showing linear regression, and "E" showing exponential regression. Error bars represent 95% confidence interval derived from Bootstrap method.

Fig. 4. Relationships among soil metal exceedance and underlying processes. (**A**) Structural Equation Modelling (SEM) of irrigation, mining, plant pumping effect, leaching, and weathering on exceedance rate and hazard level of As (n=2149, χ^2=4.45, Bootstrap P = 0.41, root mean square error of approximation (RMSEA)=0.04, standardized root mean squared residual (SRMR) = 0.009, goodness-of-fit index (GFI) = 0.999). "***" denotes significant effect with p value less than 0.001; "**" denotes significant effect with p value less than 0.01; "*" denotes significant effect with p value less than 0.05; "." denotes effect with p value less than 0.1. (**B**) SEM of Cd (n=2379, χ^2 =0.57, Bootstrap P = 0.95, RMSEA=0.00, SRMR = 0.003, GFI = 1.000). (**C**) Summed direct effect and indirect effects. The direct effect reflects the degree of standard deviation change in dependent variables with each one standard deviation change in a directly linked predictive variable, and indirect effect reflects the magnitude of associated change

via a indirect link. (**D**) Feature importance assessed by Shapley Additive Explanations (SHAP) (text S1.4.4). The larger the Shapley value, the more important a variable on the X axis is (*38*).

Science
AAAS

Supplementary Materials for

Global soil pollution by toxic metals threatens agriculture and human health

Deyi Hou*, Xiyue Jia, Liuwei Wang, Steve P. McGrath, Yong-Guan Zhu, Qing Hu, Fang-Jie Zhao, Michael S. Bank, David O'Connor, Jerome Nriagu

*Corresponding author. Email: houdeyi@tsinghua.edu.cn

The PDF file includes:

1 Materials and Methods

1.1 Dataset

1.1.1 Toxic metal species

In the present study, we selected the following seven toxic metals and metalloids (herein toxic metals): arsenic, cadmium, chromium, cobalt, copper, nickel, and lead. These toxic metals were selected because they represent important soil pollutants, as evidenced by their toxicity, widely observed exceedance, and extensive anthropogenic activities causing influx into soil ecosystems. Mercury also meets these criteria, but it is excluded from the present study because the transport mechanism of mercury differs greatly from the other heavy metal(loid)s due to the volatility of elemental mercury. Additional information and rationale for each metal is provided below:

Arsenic (As): Arsenic is a human carcinogen, and long-term exposure could lead to skin cancer, bladder cancer, and lung cancer (*52*). Rice accumulates up to 10 times more arsenic than other major food crops and is the major pathway of arsenic exposure through food (*53*). In China, arsenic contamination accounted for nearly 17% of all soil quality exceedances (*10*). Soil As was found to exceed health guidance level in 1157 of the 1867 national priority list sites identified by the USEPA (*54*). Soil arsenic maybe a factor contributing to endemic arsenicosis as was reported in Guizhou in China (*55*), Comarca Lagunera in Mexico (*56*), and Antofagasta in Chile (*57*).

Cadmium (Cd): Cadmium is one of the most mobile and bioavailable heavy metal(loid)s in soil. It can be absorbed by crop plant roots and enter grains. Cadmium is a human carcinogen, and causes damages to human kidneys, skeletal and respiratory systems (*58*). In China, cadmium accounted for 43% of all soil quality exceedance based on a national soil quality survey (*10*). The accumulation of cadmium in rice is of particular concern for Asian countries where people tend to consume large amounts of rice products (*59*). In the US, soil Cd was found to exceed health guidance level in 1011 of the 1867 national priority list sites identified by the USEPA (*54*).

Chromium (Cr): Chromium is a commonly used industrial mineral. It exists in two stable valence states: Cr(III) and Cr(VI). While Cr(III) is of low toxicity, Cr(VI) is highly toxic. Soil contamination by Cr(VI) can be caused by metal processing, tannery, steel welding, and pigment production. Environmental exposure of Cr(VI) can cause renal damage, allergy and asthma, and cancer of the respiratory tract (*60*). In the US, soil Cr was found to exceed health guidance level in 1122 of the 1867 national priority list sites identified by the USEPA (*54*).

Cobalt (Co): Cobalt is a key element of lithium-ion batteries. It is also a by-product of copper and nickel mining and smelting. Although cobalt is an essential constituent of specific vitamins, excessive intake of Co can result in hearing and visual impairment, cardiovascular and endocrine impacts (*61*). In the US, soil Co was found to exceed health guidance level in 425 of the 1867 national priority list sites identified by the USEPA (*54*).

Copper (Cu): Copper is an important constituent of many enzymes, but excessive level of copper in soils can negatively impact plant growth, and accidental exposure can cause health effects in humans (*62*). In China, copper contamination accounted for nearly 13% of all soil quality exceedance (*10*). In the US, soil Cu was found to exceed health guidance level in 926 of the 1867 national priority list sites identified by the USEPA (*54*).

Nickel (Ni): Nickel is an essential element for plant growth, required for the functioning of a number of enzymes such as urease. However, excessive nickel uptake by plants can result in phytotoxicity in plants (*63*). Nickel compounds have also been classified as a human carcinogen, and nickel exposure by sensitized individuals can result in skin allergy (*64*). In China, nickel contamination accounted for nearly 30% of all soil quality exceedances (*10*). In the US, soil Ni was found to exceed health guidance level in 860 of the 1867 national priority list sites identified by the USEPA (*54*).

Lead (Pb): Lead is one of the chemicals of greatest public health concern to the World Health Organization (*65*). Damage caused by lead exposure to child intellectual development is irreversible, with even low-level exposure linked to impaired neurological development and reduced IQ (*66*). A recent population-based cohort study in the US showed that the attributable percentage of blood Pb level (BLL) to all-cause mortality was revealed to be 18%, or an estimated 0.4 million deaths per year in the US, thus, making Pb exposure comparable to tobacco smoke as a major cause of mortality (*67*). In China, lead contamination accounted for nearly 9% of all soil quality exceedance (*10*). In the US, soil Pb was found to exceed health guidance level in 1287 of the 1867 national priority list sites identified by the USEPA (*54*).

1.1.2 Toxic metal concentrations

A systematic literature search (fig. S1) was conducted to synthesize a global database of soil toxic metal concentrations (last updated on September 16th, 2024). The following keyword combination was used: TOPIC: ("soil" OR "land" OR "geochem*") AND TOPIC: ("Spatio temporal" OR "regional scale" OR "provincial" OR "province" OR "county" OR "mapping" OR "map" OR "spatial distribution" OR "spatial variability" OR "spatial variation" OR "spatial interpolation" OR "hectare" OR "acre" OR "km" OR "principal component analysis" OR "kriging" OR "GIS" OR "multi-site" OR "multiple sites" OR "forest sites") AND TOPIC: ("metal*" OR "cadmium" OR "cd" OR "cobalt" OR "copper" OR "nickel" OR "chromium" OR "arsenic" OR “Pb” OR "soil pollut*" OR "soil contam*" OR "trace element" OR "toxic element") NOT TOPIC: ("marine" OR "ocean"). The search was conducted using the Web of Science tool, covering the following databases: Web of Science Core Collection, MEDLINE, Data Citation Index, Biosis Previews, Inspec, SciELO Citation Index, Chinese Science Citation Database, KCI-Korean Journal Database. Articles and reviews published in English were retrieved for further screening. Based on the peer-reviewed studies, a snowball method from the references of relevant papers were used to identify additional database own by pertaining government agencies. We used broad search terms to locate studies across a wide range of geographic areas, which resulted in a large number of initial search results requiring screening and paper downloading. Among others, the LUCAS topsoil dataset used in this work was made available by the European Commission through the European Soil Data Centre managed by the Joint Research Centre (JRC), http://esdac.jrc.ec.europa.eu/".

A total number of 82,530 documents were identified by the search. The first round of screening was conducted based on title and abstract. Studies meeting the following criteria were retained: related to the pertaining toxic metals, focus on regional distribution rather than a specific pollution source, and the studied area is likely to exceed 10 km^2. A total of 5,933 studies were retained during this step. Subsequently we intended to retrieve full text documents for all these studies, with success for 5218 studies. For the studies with full text obtained, toxic metal concentrations were extracted from the published literature or pertaining data source. A number of studies were further excluded due to the following reasons. Firstly, studies conducted before 2000 are excluded. This criterion allows us to minimize the impact of temporal change in toxic metal concentrations attributed to anthropogenic activities. Secondly, studies that focus on contaminated sites are excluded. This is because such studies render limited regional implications, and including them in the present study may over-estimate the spatial scale of toxic metal pollution. Nevertheless, it should be noted that this criterion also results in a limitation of our study, which is that it would underestimate the impact of such “hot spots”. Thirdly, studies are excluded if the studied area is too small (<10 km^2) or sampling was not representative of the studied region. Fourthly, studies are excluded if the data quality is questionable. Methods accepted for quantifying toxic metals are laboratory based analytical procedures with rigorous quality assurance / quality control (QA/QC), including graphite furnace Atomic Absorption Spectroscopy (AAS), Inductively Coupled Plasma-Mass Spectrometry (ICP-MS), and Inductively Coupled Plasma Optical Emission spectroscopy (ICP-OES). Some research studies used portable X-ray Fluorescence Analyzer (XRF) to quantify toxic metal concentrations; however, these results are considered to be of lower accuracy and therefore excluded from the database. The final database combines 1493 published studies, including 796,084 soil samples collected from 91 countries. The distribution of sampling sites is shown in fig. S2.

The toxic metal concentrations from various studies were synthesized using 10-km x 10-km grids in EPSG:4326 (WGS84). The 10 km resolution was chosen because it is large enough to evaluate regional-scale spatial distribution rather than site specific pollution events, and it is also small enough to reasonably capture small-scale variation due to natural background, e.g. associated with differing parent materials and weathering conditions during geologic time spam, as well as different levels of anthropogenic pollution from atmospheric deposition and agricultural practice (*6, 68*). Some studies have collected soil toxic metal concentrations for soils of various depth. In the present study, only the most surficial soil concentrations were used. Moreover, soil concentrations measured for soil interval below 30 cm were systematically removed. This is consistent with most existing regional studies on soil pollution and soil properties (*68, 69*).

The concentration data allowed us to identify predictive variables that influence the spatial distribution of toxic metals (see Section 1.1.3 and Section 1.3.1). These variables are then used to evaluate the probability of toxic metal concentration exceedance. In the present study, concentration data were binary-coded using the pertaining threshold values, with concentration lower than or equal to threshold assigned zero and concentration higher than threshold assigned one. This methodology is driven by the lack of abundant toxic metal concentration at high resolution. Preliminary modeling exercise shows that regression on toxic metal concentration has limited predictive power. Therefore, this study focuses on the prediction of toxic metal exceedance, and the pertaining thresholds are discussed in Section 1.2.

1.1.3 Predictive variables

A series of covariates were used to construct predictive models for the distribution of toxic metal exceedance. The variables were selected based on the following three criteria: 1) there is a potential causal relationship between the covariate and toxic metal concentrations, and the relationship may either be direct or indirect; 2) existing regional studies have shown that the category of variables have significant correlation with certain toxic metal concentrations; 3) there are available global database of the covariates which can be used to derive values to the spatial resolution of the present study. All variables were resampled and reprojected to match the 10 km resolution grid of the toxic metal distribution. Some variables were transformed to obtain numerical values, and some variables were re-calculated to obtain weighted average for each 10 km x 10 km cell. More details are described below.

1.1.3.1 Geological variables

The geochemical, mineralogical, and physical properties of soil parent materials, i.e. rock lithology types, play an important role in soil properties (*13*, *70*). For example, cadmium concentrations in sedimentary rocks are typically higher than igneous rocks (*17*). We collected geological covariates from a global lithological map (GLiM) composed of 13 lithological classes with a spatial resolution of 0.5 degrees (*71*). The GLiM represents the rock types of the Earth surface with 1,235,400 polygons. The 13 lithological types include: evaporites, metamorphics, acid plutonic rocks, basic plutonic rocks, intermediate plutonic rocks, pyroclastics, carbonate sedimentary rocks, mixed sedimentary rocks, siliciclastic sedimentary rocks, unconsolidated sediments, acid volcanic rocks, basic volcanic rocks, intermediate volcanic rocks. For the present study, we derived the proportion of different lithological types for each 10 km cell, subsequently we used these 13 lithological variables as predictive variables.

1.1.3.2 Climatic variables

Soil is formed from the weathering of rocks, and the weathering process is quantitatively the most important source of natural toxic metals in soil (*17*). As the weathering process is largely affected by climatic conditions, climatic variables such as temperature, precipitation, frost, and evaporation may play a critical role in determining soil toxic metal concentrations (*72*). In the present study, we collected climatic data from CRU TS v. 4.05, which was developed and improved principally by the UK's Natural Environment Research Council (NERC) and the US Department of Energy. CRU TS was generated by the interpolation of monthly climate data on $0.5° \times 0.5°$ grid (*73*). The data of 9 covariate layers, including diurnal temperature range (DTR), ground frost frequency (GFR), near-surface temperature (TMP), near-surface temperature maximum (TMX), near-surface temperature minimum (TMN), potential evapotranspiration (PET), precipitation (PRE), vapour pressure (VAP), and wet day frequency (WET) in the time period of 2001 to 2020 were used, and their values of maximum, minimum, mean, median and standard deviation across the 20 years were calculated to capture the most predictive climate variables. To account for the effects of plant pumping, we also included transpiration data into models. This dataset was developed by Zhang et al through robust diagnostic models, which covers the period of 1981 to 2012 (*74*). We calculated the aforementioned five statistical measures using data from 2001 to 2012 and incorporated them into the models.

1.1.3.3 Soil texture and basic physico-chemical properties

Soil texture such as clay content and soil physico-chemical properties such as soil organic content have been widely used as co-variates to predict the regional distribution of toxic metals in soil (*24*, *25*, *75-78*). We collected data regarding soil texture and basic physico-chemical properties from Harmonized World Soil Database (HWSD) v 1.2 (*79*). HWSD is a result of the joint efforts of Food and Agriculture Organization of the United Nations (FAO), the International Institute for Applied Systems Analysis, ISRIC-World Soil Information, Institute of Soil Science, Chinese Academy of Sciences and Joint Research Centre of the European Commission. It contains over 15 000 different soil mapping units and the layer has a spatial resolution of 30 arc-seconds. In the present study, we used 4 variables for soil texture: topsoil gravel content, topsoil sand fraction, topsoil silt fraction, and topsoil clay fraction. For the basic physicochemical properties, 12 variables were used: topsoil bulk density, reference bulk density, topsoil organic carbon, topsoil pH, topsoil CEC (clay), topsoil CEC (soil), topsoil base saturation, topsoil TEB, topsoil calcium carbonate, topsoil gypsum, topsoil sodicity, topsoil salinity.

1.1.3.4 Topography

Land topography affects soil-forming rock weathering processes, and it also affects how surface runoff accumulates and infiltrates (*40*). It would influence how heavy metal and metalloid elements are leached out of rock/soil in one place, and then adsorbed and accumulated in soil at another place (*17, 80*). Therefore, topographic parameters may be used as predictors for regional soil pollution by toxic metals (*23, 81*). In the present study, we collected elevation and slope data from the Global Terrain Slope and Aspect Dataset (*82*). The dataset has a resolution of 5 minutes. In this dataset, slope gradient is divided into 8 types: 0 % ≤ slope ≤ 0.5 %, 0.5 % < slope ≤ 2 %, 2 % < slope ≤ 5 %, 5 % < slope ≤ 10 %, 10 % < slope ≤ 15 %, 15 % < slope ≤ 30 %, 30 % < slope ≤ 45 %, Slope > 45 %. The numerical value is expressed as the percentage of each slope type times 1000.

1.1.3.5 Socioeconomic variables

Socioeconomic indicators, especially those related to agricultural and industrial production, are important predictors of soil pollution (*6, 25, 68*). Anthropogenic sources account for the majority of atmospheric emission of various toxic metals (*13, 15*), which is a main contributor of regional soil pollution. While soil pollution may intensify when population density and industrial output grow, the input of toxic metal may also decrease when countries become more developed and environmental governance strengthen. In Europe, the industrial input of cadmium peaked in the 1960s and has decreased since then (*17*). To capture the complex dynamics revolving around anthropogenic activities, we collected a series of variables associated with economic and social development for model building. The gross domestic product (GDP) and population density data were obtained from the Socioeconomic Data and Applications Center initiated by NASA. GDP data contain two layers for 1990 and 2025 with a special resolution of a 15 × 15 minute grid (*83*). For population density, we used the Gridded Population of the World (GPWv4), Version 4: Population Density Adjusted to Match 2015 Revision UN WPP Country Totals, Revision 11 for 2000, 2005, 2010, 2015 and 2020, at 2.5 arc-minute resolution (*84*). We also collected 5 country-level variables from the World Bank: ore/metal exports, mineral rents, mineral depletion, ores/metals imports and mortality caused by road traffic injury. The covariate, adjusted savings for mineral depletion, were normalized by the area of country. All the covariates were resampled to a 10 km × 10 km grid.

Land use type is a proxy of various anthropogenic activities which correlates with toxic metal input processes, and maybe used as a co-variate of soil pollution (*25, 85-87*). Land cover data were collected from the Land Cover CCI Climate Research Data Package (CRDP) provided by the European Space Agency (ESA). These data packages contain the annual land-use map from 1992 to 2015 at a 300-meter resolution. We selected the data from 2015 to construct a predictive variable for land use. The land cover is divided into 22 types in the original data, and we reorganized it to construct 5 land cover related variables, including agricultural land use, bare areas, forest land use, settlement land use, and other vegetation cover. We calculated the percentage of different land-cover types within the 10-kilometer grid.

Many existing studies have shown that both inorganic and organic fertilizer production and application are important sources of soil heavy metals (*17, 21*). Therefore, we used nitrogen fertilizer application, nitrogen in manure production, phosphorus fertilizer application, and phosphorus in manure production data derived from Global Fertilizer and Manure, Version 1 Data Collection, to quantify the influence of fertilizer on soil contamination (*88*). The dataset has a special resolution of 0.5 degrees. Agricultural land irrigation is also indicative of the intensity of agricultural activity, and irrigation with contaminated runoff can cause heavy metals accumulation in soil. We used the percentage of irrigated area, the percentage of irrigated area with groundwater, and the percentage of irrigated area with surface water from the Global Map of Irrigation Areas Version 5, developed by FAO as predictive variables (*89*). The resolution of these raters was 5 minutes.

Atmospheric deposition is known as an important source of heavy metals in soil on a regional scale (*90, 91*). We used the global power plant emission database developed by Tong et al, which includes CO_2 and air pollutant emissions (SO_2, NO_x and primary PM2.5) from the main power plant in 231 countries or regions (*92*). The four covariate grids are at a special resolution of 0.1 degrees. These variables do not directly represent toxic metal deposition; however, they may serve as a proxy of toxic metal deposition.

1.2 Regulatory thresholds

Regulatory thresholds were obtained from 11 countries, including Austria (*93*), Belgium (*93*), Canada (*94*), China (*95, 96*), Denmark (*93*), Finland (*93*), France (*93*), Germany (*93*), Italy (*93*), Netherland (*93*), and the United States (*97*). We have included both screening values, which usually trigger site-specific health risk assessment, and intervention values, which usually mandates cleanup efforts. Table S1 summarize these threshold values. The regulatory thresholds vary by orders of magnitude in different countries, for different land use, and under different soil conditions (Table S2). In the present study, we intend to be moderately conservative, and we have selected the 25 percentile values for the purpose of the modeling. A separate set of thresholds were derived for agricultural soils, as a smaller number of countries have such thresholds available. As Table S1 shows, the agricultural thresholds tend to be similar or lower than the thresholds for human and ecological health, and Cd has the most significant difference, i.e. 6 mg/kg for human health and ecological threshold versus 1 mg/kg for agricultural threshold. Soil concentration data were converted to binary data based on the selected regulatory thresholds, using the following Inference method.

1.3 Exceedance inference

Here we consider that each 10 km x 10 km grid consists of many smaller grids, and whether the true toxic metal concentration in each smaller grid exceeds the above threshold follows a Bernoulli distribution, which is a common discrete distribution categorized as “exceed” or “not exceed” (*98*). When deriving whether toxic metal concentration in a target area exceeds a threshold, existing large-scale studies have mainly used three different treatment methods. Firstly, some studies used the maximum concentration in each pixel to derive whether it exceeds the threshold (*99*), which renders a conservative and very likely overestimate of exceedance. Secondly, some other studies used the arithmetic or geometric mean in each pixel to derive whether it exceeds the threshold (*11, 14*). Thirdly, some studies used a simple proportion of samples exceeding threshold to represent the probability of exceedance (*32, 100*). In the present study, we used arithmetic mean to represent each pixel because: 1) it would reduce the likelihood of overestimating exceedance in comparison with using maximum concentrations; 2) it better represents the “average” exposure scenario than a geometric mean; and 3) it reduces the likelihood of overestimating exceedance rate with the simple proportion of sample exceedance (see detailed inference below and Table S8). As each 10 km x 10 km grid maybe covered by different studies, a synthesis procedure has been employed to integrate data from various sources, using the following equations:

$$c_{i,j} = \frac{\sum_l c_{l,j} \cdot \frac{A_i}{A_l} \cdot n_{l,j}}{\sum_l \frac{A_i}{A_l} \cdot n_{l,j}} \tag{1}$$

where $c_{i,j}$ is the average toxic metal concentration in grid *i* for toxic metal *j*; $c_{l,j}$ is the average toxic metal concentration in study *l* for toxic metal *j*; A_i denotes the size of the land area in the i_{th} cell covered by the l_{th} study; A_l denotes the total area covered by the l_{th} study; $n_{l,j}$ is the total sampling points in the *l*th study for the toxic metal *j*. The standard deviation of toxic metal concentrations in each grid was derived using the following equation:

$$std_{i,j} = \frac{\sqrt{\sum_l \left(std_{l,j} \cdot \frac{A_i}{A_l} \cdot n_{l,j}\right)^2}}{\sum_l \frac{A_i}{A_l} \cdot n_{l,j}} \tag{2}$$

where $std_{i,j}$ is the standard deviation of toxic metal concentration in grid *i* for toxic metal *j*; $std_{l,j}$ is the standard deviation of toxic metal concentration in study *l* for toxic metal *j*. For grids with access to individual toxic metal concentrations at each sampling point, or studies covering no more than one grid, the above equations were directly used. For studies covering a large area, to avoid overestimating model accuracy owing to spatial autocorrelation, only 30% of the grids in each study area were randomly selected for modeling. Moreover, the exceedance state was derived for each randomly selected grid based on the following inference procedure. According to preliminary analysis of our dataset as well as previous studies, soil toxic metal concentrations tend to follow positively skewed distribution, often approximating lognormal distribution, $Lognormal(\mu_{i,j}, \sigma_{i,j})$, which is also why some existing studies used geometric mean rather than arithmetic mean to derive exceedance rate. The parameters of the log-normal distribution, $\mu_{i,j}$ and $\sigma_{i,j}$ can be estimated based on the following equations:

$$\mu_{i,j} = \ln(c_{i,j}) - \frac{1}{2}\left(\ln\left(\frac{{std_{i,j}}^2}{{c_{i,j}}^2} + 1\right)\right) \tag{3}$$

$$\sigma_{i,j} = \sqrt{\ln\left(\frac{{std_{i,j}}^2}{{c_{i,j}}^2} + 1\right)} \tag{4}$$

Random sampling of each grid was then conducted using the above lognormal distribution, and repeated for $\frac{A_i}{A_l} \cdot n_{l,j}$ times. Based on a preliminary analysis of the robustness of deriving exceedance probability with sampling data, a cut-off value of 15 sampling points was selected to decide whether only in-grid data were used or out-of-grid data were also used. For grids with less than 15 sampling points, out-of-grid search was conducted to locate the closest sampling points until total sampling points reach 15 or reach a 1^o by 1^o range. Weighted average was used to ascertain the concentration of toxic metal in the target grid. According to inverse distance weighting interpolation, we have assigned weights as the reciprocal of the distance from the point to the center of the grid (*101*). For in-grid data, we assumed their distance from the grid's center to be half the side length of grids owing to their congruent importance in determining the toxic metal's level in the target grid. The selected sampling results were then used to derive the exceedance state for the corresponding grid. Finally, we obtained 30,122 data points for As, 31,138 data points for Cd, 25,909 data points for Co, 31,026 data points for Cr, 31,500 data points for Cu, 30,509 data points for Ni and 31,792 data points for Pb.

We also conducted the following inference and analyses to compare differences among simple proportion of sample exceedance, the aggregated probability of exceedance within each grid, and the proportion of grids with average concentrations exceeding thresholds. The inference procedures of the aggregated probability of exceedance are as follows. The probability of whether the toxic metal concentration in a randomly selected 10 km x 10 km grid exceeds a threshold may also be assumed to equal to the percentage of smaller grids that are in "exceed" state. The probability of exceedance rate should follow Beta distribution, $Beta(\alpha_{i,j}, \beta_{i,j})$ (*98*), and can be described by using the following equation:

$$f_{P_{i,j}}(p) = \frac{1}{B(\alpha_{i,j}, \beta_{i,j})} \, p^{\alpha_{i,j}-1}(1-p)^{\beta_{i,j}-1} \tag{5}$$

where p is the probability of toxic metal exceedance; $f_{P_{i,j}}(p)$ is the probability density function of toxic metal exceedance rate in the *i*th grid for the *j*th toxic metal; $\alpha_{i,j}$ is the shape parameter corresponding to smaller grids exceeding threshold, and represented by known sampling points in the larger grid exceeding threshold; $\beta_{i,j}$ is the shape parameter corresponding to smaller grids not exceeding threshold, and represented by known sampling points in the larger grid not exceeding threshold. $B(\alpha_{i,j}, \beta_{i,j})$ is defined by the following equation:

$$B(\alpha_{i,j}, \beta_{i,j}) = \frac{\Gamma(\alpha_{i,j})\Gamma(\beta_{i,j})}{\Gamma(\alpha_{i,j} + \beta_{i,j})} \tag{6}$$

where Γ denotes Gamma function. When in-grid sample number exceeds 15, we can directly infer the probability of toxic metal exceedance using Equation 7:

$$P_{i,j} = P(p \geq p_j) = \int_{p_j}^{1} \frac{1}{B(\alpha_{i,j}, \beta_{i,j})} \, p^{\alpha_{i,j}-1}(1-p)^{\beta_{i,j}-1} \, dp \tag{7}$$

where $P_{i,j}$ is the probability of toxic metal's exceedance rate over exceedance rate threshold p_j. In this study, p_j is 0.5 and when $P_{i,j}$ exceeds 0.5, we consider the grid is in "exceed" state. For grids with less than 15 samples, we employed the following Bayesian Inference procedure (*102, 103*). The following search algorithm was used to include out-of-grid data. We started from grid i and gradually increased search radius r until the number of sampling points in the generated search area H reach 15. The probability of toxic metal exceedance rate follows Beta distribution $Beta(\alpha_{H,j}, \beta_{H,j})$, where $\alpha_{H,j}$ is sampling points in area H exceeding threshold; $\beta_{H,j}$ is sampling points in area H not exceeding threshold. According to Tobler's First Law of Geography, near things are more related than distant things (*104*); therefore, this Beta distribution in the larger area H represents prior information for the probability distribution of toxic metal exceedance rate in the smaller 10 km by 10 km grid which holds close proximity to area H. On the other hand, the number of sampling points exceeding toxic metal threshold in the 10 km by 10 km grid follows $Binomial(n_{i,j}, p_{i,j})$ distribution (*98*), and can be described by using the following equation:

$$f(k, n_{i,j}, p_{i,j}) = \Pr(X = k) = \binom{n_{i,j}}{k} \, {p_{i,j}}^{k}(1-p_{i,j})^{n_{i,j}-k} \tag{8}$$

where $f(k, n, p)$ is the probability density function of the number of sampling points exceeding toxic metal threshold; $n_{i,j}$ is the number of sampling points in grid *i* for toxic metal *j*; *k* is the number of sampling points exceeding toxic metal threshold; $p_{i,j}$ is the toxic metal exceedance rate. Bayes theorem enables us to infer the posterior distribution based on experimental data and prior distribution (*105*):

$$f\left(p_{i,j}\middle|\mathcal{D}_{i,j}\right) \propto \mathcal{L}\left(\mathcal{D}_{i,j}\middle|p_{i,j}\right)g\left(p_{i,j}\right) \tag{9}$$

where $f(p_{i,j}|\mathcal{D}_{i,j})$ is the posterior distribution, $\mathcal{L}(\mathcal{D}_{i,j}|p_{i,j})$ is the likelihood function, $g(p_{i,j})$ is the prior distribution and $\mathcal{D}_{i,j}$ represents experimental data which is actually the data in grid i. Based on Tobler's Law the prior distribution of toxic metal exceedance rate in grid i approximates the distribution of exceedance rate in area H, namely $g(p_{i,j}) = Beta(\alpha_{H,j}, \beta_{H,j})$. The likelihood function is given by the binomial distribution, namely $\mathcal{L}\left(\mathcal{D}_{i,j}\middle|p_{i,j}\right) = Binomial(n_{i,j}, p_{i,j})$. Given that Beta distribution is a conjugate prior for Binomial distribution, we can get the posterior distribution of $p_{i,j}$ given $\mathcal{D}_{i,j}$ as described by the following equation (*106*) .

$$f\left(p_{i,j}\middle|\mathcal{D}_{i,j}\right) = \frac{\mathcal{L}\left(\mathcal{D}_{i,j}\middle|p_{i,j}\right)g\left(p_{i,j}\right)}{\int \mathcal{L}\left(\mathcal{D}_{i,j}\middle|p_{i,j}\right)g\left(p_{i,j}\right)dp_{i,j}} = Beta\left(\alpha_{H,j} + \alpha_{i,j}, \beta_{H,j} + \beta_{i,j}\right) \tag{10}$$

Then, the derived Beta distribution is used to inference whether grid i is in in "exceed" state for toxic metal j based on Equation 7. Table S8 provides a comparison of these calculation results. It should be noted that the present study uses concentrations from regional studies which did not distinguish agricultural land from non-agricultural land. To assess errors that may be introduced, we used data from an EU-wide study to compare these two rates for different land use under the same thresholds. It was found that metal exceedance rate for agricultural land only increased slightly in comparison with that for all land uses, for both low and high threshold values (table S7), confirming the validity of this method.

1.4 Modeling methods

First, the entire dataset was randomly split into two subsets: 80% of the data was used as a training dataset to calibrate the model, and 20% of the data was used as an evaluation dataset to assess how well the calibrated model predicts. The general distributions of each soil toxic metal exceedance were similar in the training set and test set. We conducted preliminary experiments to explore the performance of ten machine learning models in predicting toxic metal's exceedance based on all predictive variables, including extremely randomized trees (ERT), random forest (RF), Adaptive Boosting, Gradient Boosting, eXtreme Gradient Boosting, Support Vector Machine, Multi-layer Perceptron, K-Nearest Neighbors, Decision Tree and Logistic Regression with L2 regularization. Among these machine learning algorithms, the accuracy of ERT was the highest for all toxic metals. ERT is a decision tree-based ensemble method that is similar to RF but uses a different technique to build the individual trees (*27*). In an ERT model, a large number of decision trees are grown using random subsets of features, and the final prediction is made by aggregating the predictions of all the trees in the ensemble. Compared to RF, ERT introduces additional randomness in the tree-building process by using random splits for each node in the tree (*107*). Specifically, for each node in the tree, a random subset of features is selected and a random threshold is chosen for each feature to split the data. ERT is known to render robust and satisfying performance in classification for nonlinear issues and imbalanced dataset with faster training speed, and it has been widely used in a variety of research areas (*108, 109*). Based on assessment results from the preliminary experiments, ERT was selected as the optimal model to quantify the high-dimensional nonlinear relationship between toxic metals' exceedance and the wide ranges of predictive variables.

1.4.1 Feature selection

Feature selection plays a critical role in model development, which aims to drop out redundant variables, and thus to avoid overfitting and multicollinearity, improve model performance and interpretability, as well as to reduce computational costs (*110*). In the present study, we conducted a two-step method to ensure our final feature set did not involve redundant information (e.g multicollinearity). For the first step, recursive feature elimination (RFE) was conducted to select features from a collection of 116 predictive variables. RFE is a widely used method for feature selection, which iteratively removes the weakest variables according to the importance of features (*111*). Feature importance was evaluated by mean decrease in node impurity (MDI) via Gini index in this study, and features were removed iteratively until only one remained. The importance of features in the model can be determined according to the order in which variables were eliminated. The later they are removed from the model, the more important the feature is. Then, a feature set with the least number of features and highest accuracy in the model was selected out primarily. However, RFE can remove unimportant variables, but it cannot remove important variables with strong collinearity. Therefore, the second step was conducted to further eliminate redundant features. Pearson correlation coefficient (r) was calculated to provide us an insight of the collinearity strength among variables. Features with high correlation with others ($r \geq 0.5$) are identified as collinear variables. For collinear variables, we leave only the most important variables indicated by feature importance to avoid multicollinearity in models. Finally, for As, there were 25 features remaining in AT models, and 20 features remained for Cd, 40 for Co, 18 for Cr, 13 for Cu, 10 for Ni, and 19 for Pb. The final selected features for each metal are shown in Attachment 2 of (*38*).

1.4.2 Hyperparameter tuning

Hyperparameter tuning was conducted following feature section. Five hyperparameters were optimized with grid search, including the number of trees, the maximum depth of the tree, the minimum number of samples required to be at a leaf node, the minimum number of samples required to split an internal node and the function to measure the quality of a split. Ranges of hyperparameters were carefully set to maximize model accuracy and avoid overfitting (see Table S3). As the dataset was imbalanced, a parameter was set to automatically balance sample weight for every tree grown, meaning minority class was assigned higher weight in the process of model development. During hyperparameter tuning, 5-fold cross-validation was conducted and F1-score (see section 1.4.3 for definition), an effective metric for imbalanced dataset, was used to assess model performance. The optimal hyperparameter settings for different models were listed in Table S4. Models used to predict the toxic metal exceedance in agricultural land were trained with similar procedures.

1.4.3 Model evaluation

A group of scoring metrics were adopted to assess the performance of the calibrated models, including balanced accuracy (BA), sensitivity, specificity, F1 score, average precision (AP), the area under the Receiver Operating Characteristic Curve (AUC) and Cohen's kappa coefficient (KIA). The prediction results were divided into 4 types in terms of true positive (TP), false negative (FN), true negative (TN) and false positive (FP). Balanced accuracy was used to evaluate the model performance of classification, which is particularly useful when the input data is imbalanced.

$$Balanced\ accuracy = \frac{Sensitivity + Specificity}{2} \tag{11}$$

Sensitivity and specificity are the true positive and negative rates, respectively, and sensitivity is also known as recall.

$$Sensitivity = \frac{TP}{TP + FN} \quad (12)$$

$$Specificity = \frac{TN}{TN + FP} \quad (13)$$

F1 score is the harmonic mean of the precision and recall. Recall measures the proportion of true positives that are correctly identified by the models, while precision measures the proportion of identified positives that are actually positives. F1 score is regarded as an effective metric in evaluating model performance trained from imbalanced data. Therefore, it was not only used to evaluate the final model performance, but also used in feature selection and hyperparameter tuning. The closer the F1 score is to 1, the better the prediction performance of the model.

$$F_1 = 2 \times \frac{recall \times precision}{recall + precision} = \frac{2 \times TP}{2 \times TP + FP + FN} \quad (14)$$

AP is the area under precision-recall curve (see fig. S3). By taking into account both precision and recall, AP provides a more informative and reliable measure of performance than many other metrics that only consider one aspect of the model's accuracy. AUC is the area under the Receiver Operating Characteristic Curve. It measures the overall quality of the model's predictions by quantifying the trade-off between the true positive rate (sensitivity) and the false positive rate (1-specificity) at various classification thresholds. Cohen's kappa is a common metric used to evaluate the agreement between the measured values and predicted results. When Cohen's kappa is higher than 0.8, it indicates that the model performance is excellent (*112*).

$$\mathrm{KIA} = \frac{P_{obs} - P_{exp}}{1 - P_{exp}} \quad (15)$$

$$P_{obs} = \frac{TP + TN}{N} \quad (16)$$

$$P_{exp} = \frac{(TP + FN)(TP + FP) + (FP + TN)(FN + TN)}{N^2} \quad (17)$$

where N is the number of samples in test set.

Model performance for different toxic metals derived from ERT is shown in Table S6. Although the datasets are imbalanced for these toxic metals, where positive samples account for less than 7% of the whole dataset, models present predictions with high accuracy for both positive and negative samples on the test dataset (20% of the data, which was randomly selected while maintaining the relative distribution of high and low values). The sensitivity of Cd-AT model is higher than 0.8, and the specificity of all seven toxic metals is closes to 1. The extremely imbalanced distribution of positive and negative samples may attribute to the relatively low sensitivity values for Cd in HHET model. Apart from sensitivity and specificity, comprehensive metrics also indicate that our models are well-trained. Co-AT obtains the highest KIA as 0.86, followed by Ni-AT (0.78), suggesting that the model performance is excellent. The KIA of other toxic metals for both AT and HHET models is higher than 0.6, showing that the models for these toxic metals are good. BA, F1-score and AP showed congruent patterns with KIA (Table S6).

Data imbalance often hinders model training. In order to make precise and robust predictions, we have taken several measures to reduce the impact of unbalanced data sets on the model, including: (1) selecting ERTs, which is one of the most suitable algorithms for imbalanced data; (2) adjusting class weight inversely proportional to sample distributions in model training, which give more weight to positive samples; (3) employing F1-sore as scoring metrics in feature selection and parameter tuning, which provides a balanced measure of performance that takes into account both false positives and false negatives.

We used the best models to generate the probability of being polluted for 2,000,000 pixels and developed the global pollution probability maps for different toxic metals. We then excluded the pixels covered by desert and permafrost, with 1,290,000 pixels remained finally. The map of permafrost was obtained from National Science Foundation Arctic Systems Science Program (*113*). The dataset of desert is at 1:10 million scale, which was derived originally from the Florida Resources and Environmental Analysis Center's Physical Map of the World and held by Stanford currently (*114*). We then display the area affected by all toxic metals by calculating the maximum probability of all toxic metal exceedance in each pixel.

1.4.4 Feature importance

The importance of covariates used to predict the probability of toxic metal exceedance was estimated by Shapley Additive Explanations (SHAP). SHAP was developed based on the game theoretically optimal Shapley Value (SV) by Lundberg and Lee (*115*). Originally, SV provides a strategy to quantify the contributions of players to the total payout. In machine learning, players can be the covariates engaged in prediction models, and payout is the prediction value. SHAP is the average marginal contribution of an evaluated feature across all coalitions of other features. The basic idea underlying SHAP feature importance is that an important feature has a larger absolute SV. The importance is measured by the average of the absolute SHAP value of the feature across the data. The larger the value of SHAP, the more important the variable. Moreover, SHAP values provide us with insights of how a given feature affects probabilities of metal exceedance (*116*), and the results are displayed in supplementary data (*38*).

1.5 Population at risk

To estimate the affected population, we need to determine a probability cutoff to classify whether the grid is exposed to high or low levels of metals in soil. In this study, we used the cutoff which makes the predicted metal exceedance equal to observed metal exceedance (*117*). The cutoffs for different metals are displayed in Table S5. The area of affected land was calculated with the following equations.

$$AL_{i,j} = \begin{cases} A_i, Prob_{i,j} \geq cutoff \\ 0, Prob_{i,j} < cutoff \end{cases} \quad (18)$$

where $AL_{i,j}$ is the area of affected land in the *i*th grid for the *j*th metal; $Prob_{i,j}$ is the probability of the *j*th toxic metal's exceedance in grid *i*.

$$AL_j = \sum_i AL_{i,j} \quad (19)$$

where $\boldsymbol{AL_j}$ denotes total area of affected land for toxic metal *j*.

The number of affected population was derived based on the following equations.

$$AP_{i,j} = \begin{cases} Popu_i, Prob_{i,j} \geq cutoff \\ 0, Prob_{i,j} < cutoff \end{cases} \quad (20)$$

where $AP_{i,j}$ is the number of affected population in the *i*th grid for the *j*th metal; $Popu_i$ denotes the number of population in 2020 in grid i , which was extracted from a dataset shared by Socioeconomic Data and Applications Center initiated by NASA (*118*).

$$AP_j = \sum_i AP_{i,j} \quad (21)$$

where $\boldsymbol{AP_j}$ denotes the number of affected population for toxic metal *j* in the world.

1.6 Agricultural land at risk

The probabilities of metals' exceedance in agricultural land are shown in Fig. 1A. The area of affected agricultural land by toxic metals was calculated by using the following equations.

$$AA_{i,j} = \begin{cases} A_i \times RA_i \times Prob_{i,j}, Prob_{i,j} \geq cutoff \\ 0, Prob_{i,j} < cutoff \end{cases} \quad (22)$$

where $AA_{i,j}$ is the area of affected agricultural land in the *i*th grid for the *j*th metal; RA_i is the ratio of agricultural land in grid i, which was derived from CRDP as introduced in Section 1.1.3.5; A_i is the area of grid i. The probability cut-off for determining whether a grid is at high risk or low risk is presented in Table S5.

$$AA_j = \sum_i AA_{i,j} \quad (23)$$

where $\boldsymbol{AA_j}$ denotes the number of affected agricultural land for toxic metal *j* in the world.

1.7 Uncertainty analysis

Several analyses were conducted to account for model uncertainties, including processes involved in data generation, feature selection, and model construction. Firstly, the entire dataset was randomly split into training and validation datasets, which renders uncertainty because different realizations of this process would result in different models. To analyze this uncertainty, we conducted a stratified bootstrap procedure for each toxic metal. In stratified bootstrapping, the subsets are constructed according to the proportion of each class, which helps to avoid the bias caused by resampling (*119*). For each metal, we performed 100 rounds of bootstrapping. The generated subsets were used to select features, build models, and estimate the probability of exceedance. Uncertainty was also introduced when we inferred whether toxic metals exceed thresholds in any specific 10 km x 10 km grid based on toxic metal concentration distribution in regional studies. Moreover, only 30% of grids were randomly selected for model development to minimize the impact of spatial autocorrelation on models. To account for these uncertainties, we generated 100 datasets with the same random procedure and build models to quantify the above uncertainties. We used the 200 model to general 95% confidence interval and calculate label stability (LS) to display the overall uncertainty mentioned above (Equation 24). The results of label stability are shown in fig. S18-19.

$$LS_{i,j} = \frac{\left|L_{0,i,j} - L_{1,i,j}\right|}{200} \quad (24)$$

Where $LS_{i,j}$ is the label stability of grid *i* for toxic metal *j*; $L_{0,i,j}$ denotes the number of models infer that the *i*th grid is in a "not exceed" state for the *j*th toxic metal; $L_{1,i,j}$ denotes the number of models infer that the *i*th grid is in a "exceed" state for the *j*th toxic metal.

Extrapolation and upscaling are also sources of uncertainty because the relationship between predictive variables and dependent variables may no longer hold true outside of the range of the training dataset. To address this uncertainty, we assessed the extent of extrapolation in our models for the 1,290,000 cells across all the involved predictive variables for each metal. The maximum and minimum values of each variable were calculated in the sampling cell, and an interpolation range was generated for the variable. Then, the proportion of variables with values falling into the interpolation range across the 1,290,000 cells was calculated to indicate the extent interpolation for each cell. This map can also reveal the representativeness of our samples. The results show that, for all metal, our samples covered most of the conditions. For all metals, 95% of cells have 90% of predictive variables inside the interpolation range (gis. S20-21). Mapping the extent of extrapolation highlighted that our dataset covered most environmental conditions, with the least represented pixels and highest proportion of extrapolation in the Southeast Asia, Russia, the central and eastern Africa, and the northern part of South America.

Apart from the uncertainties quantified above, there are several processes that would introduce additional uncertainty in this study, which was difficult to quantify but still warrant attention. Firstly, samples are not evenly distributed and the samples in regions such as northern North America, northern Asia, and Africa are relatively limited, leading to increasing uncertainties in prediction results. Secondly, there is some survivor bias effect in our dataset. The data we used are selected from studies conducted on a regional scale, to avoid overestimation of metals' exceedance caused by studies on pollution sources. However, researchers may tend to choose areas with naturally high metals' concentration, and areas suffering from intensive industrial and agricultural activities. This causes our dataset to contain more regions with metals' exceedance, which on the one hand reduced data imbalance, but on the other hand over-represented such regions. Thirdly, the spatial resolution of some predictive variables is low and most of these variables are predicted by models based on limited observed data, and these factors would also produce uncertainties.

In this study, we used the soil sampling data collected at a time period of the past two decades. The input and output of toxic metals on an annual basis are usually much smaller than toxic metal stock in soil (*17*); however, the temporal change over decadal time scales is more uncertain. In Europe, archived samples from experimental stations in the UK, France, and Denmark showed that cadmium concentration increased by 1.3~2.6 times during the 19th and 20th century (*120*), suggesting an extremely slow rate of change on decadal time scale (4.3%~6.6% per decade). However, it should be noted that serious pollution and episodic events can occur over short temporal scales and cause rapid increases in toxic metal concentrations at local sites. Here, we focus on the regional average concentration and exclude contaminated sites; therefore, the use of toxic metal data over two decades should have limited impact on the robustness of our model.

1.8 Statistical analysis

We employed structural equation modelling (SEM) to elucidate the underlying causal pathways of a variety of factors (e.g climate factors, soil properties and socioeconomical indicators) influencing the distribution and exceedance of metals in soil. SEM is a widely-used statistical approach that integrates factor analysis and regression analysis enabling the simultaneous estimation of multiple complex relationships among variables and the testing of theoretical models (*121*). To enhance the conciseness and interpretability of our model, we selected several influential variables indicated through importance analysis and constructed five indexes to characterize the drivers and processes involved in the accumulation and transportation of metals. These five indexes are weathering, leaching, plant pumping, irrigation and mining. The weathering index is derived from the diurnal temperature range, precipitation, and clay content. The irrigation index comprises the percentage of area under actual irrigation and the percentage of area irrigated with surface water. The mining index includes mineral rents (% of GDP), exports of ores and metals, and imports of ores and metals. The leaching index is represented by wet day frequency, while the plant pumping index is indicated by potential evaporation. Prior to index construction, we standardized the selected variables using the following equation to scale them within the range of 0 to 1 and eliminate the influence of extreme values.

$$x'_{i,j} = \begin{cases} \dfrac{x_{i,j}}{\bar{x}_j + 3 \times \sigma_i}, x_{i,j} \leq \bar{x}_j + 3 \times \sigma_j \\ 1, x_{i,j} > \bar{x}_j + 3 \times \sigma_j \end{cases} \tag{25}$$

where $x'_{i,j}$ represents the standardized value of variable j and observation i, and $x_{i,j}$ denotes original value. $\bar{x}_j$ is the average of variable j and σ_j represents the standard deviation value of variable j. Then, the index is formed by adding the above standardized indicators and dividing it by the number of indicators.

Apart from exceedance rate, we also explored how these factors influence hazardous levels. Hazardous level was examined by hazard quotient and hazard index, which are widely employed in health risk assessments developed by the United States Environmental Protection Agency (USEPA) (Equation 26-27) (*122*). In this process, risks associated with dermal contact, ingestion, and inhalation exposure pathways were all taken into consideration.

$$HQ = \frac{CDI}{RfD} \tag{26}$$

$$HI_{i,j} = HQ_{ing,i,j} + HQ_{inh,i,j} + HQ_{der,i,j} = \frac{CDI_{ing,i,j}}{RfD_{ing,j}} + \frac{CDI_{inh,i,j}}{RfD_{inh,j}} + \frac{CDI_{deri,j}}{RfD_{ing,j} \times ABS_{GI,j}} \tag{27}$$

where ing, inh and der represent the pathway of ingestion, inhalation and dermal contact, respectively. $HQ_{i,j}$ refers to the Hazard quotient for observation i and metal j. HI stands for hazard index. CDI denotes chronic daily intake values, which are calculated by Equation 28 to 30. RfD is reference doses, RfC denotes reference concentration, and ABS_{GI} is gastrointestinal adsorption factor. Values of these parameters used in this study are presented in Table S10.

$$\mathrm{CDI}_{\mathrm{ing},i,j} = \mathrm{C}_{\mathrm{soil},i,j} \times \frac{\mathrm{IngR} \times \mathrm{EF} \times \mathrm{ED}}{\mathrm{BW} \times \mathrm{AT}} \times \mathrm{CF} \tag{28}$$

$$CDI_{inh,i,j} = C_{soil,i,j} \times \frac{ET \times EF \times ED}{PEF \times AT} \times \frac{1\ day}{24\ hours} \tag{29}$$

$$\mathrm{CDI}_{\mathrm{der},i,j} = \mathrm{C}_{\mathrm{soil},i,j} \times \frac{\mathrm{SA} \times \mathrm{AF} \times \mathrm{ABS} \times \mathrm{EF} \times \mathrm{ED}}{\mathrm{BW} \times \mathrm{AT}} \times \mathrm{CF} \tag{30}$$

where $\mathrm{C_{soil}}$ is the concentration of metal in soil. The description and value used in this study of other parameters in Equation 28 to 30 can be found in Table S9. Hazardous level was log-transformed to achieve normality.

Before developing SEM, we initially assessed bivariate relationships among weathering, leaching, plant pumping, mining, irrigation, exceedance rate and hazardous level. We also calculated exceedance rates for various regions and different ranges of a given variable to explore the relationship among the various underlying processes that govern the accumulation of metal in soil. Based the exploratory analysis and existing theories on geological cycling of metals (*8*, *12-14*, *17*, *22*), the most complete priori models were built. The pathways between variables that did not contribute substantial information were eliminated from the priori models. The final model was selected using Akaike information criterion. Since some residuals in the data did not strictly follow a normal distribution, we conducted the Bollen-Stine bootstrap test to ascertain the significance of the final model (a good fit is indicated by Bootstrap $P > 0.10$) (*123*). To provide a comprehensive evaluation of the models' performance, we also employed other three commonly used indicators: standardized root mean squared residual (SRMR < 0.08 represents a qualified model), root mean square error of approximation (RMSEA < 0.05 stands for a good fit) and goodness-of-fit index (GIF > 0.95 for satisfactory performance) (*124*). Another crucial capacity of SEM is the examination of both direct and indirect effects between variables of interest. To comprehensively interpret our final model, we calculated the direct and indirect effects of weathering, leaching, plant pumping, mining, irrigation on the exceedance rate and hazardous level through standardized path coefficient.

2 Supplementary Discussion

Previous studies in China and Europe reported higher exceedance rates (table S7), due to lower thresholds used in those studies *(32, 100)*. To further compare a scenario with the same threshold values, we derived exceedance rate with data extracted from the EU study, which was created with a regression kriging method. Our exceedance rate estimates were slightly higher than those from the EU study, e.g. 1.4% and 4.2% from the EU study versus 2.9% and 5.1% from the present study. We attribute the discrepancy to the nature of the kriging method, which relies on spatial stationarity and is incapable of estimating robust variograms in the presence of extremely high values (*125*). We conducted supplementary analyses and found that the commonly used simple proportion method tends to yield high exceedance rates (table S7). Our machine learning results fall in between and may provide the best representation of local risks at a 10 km by 10 km grid spatial resolution.

3 Supplementary Figures

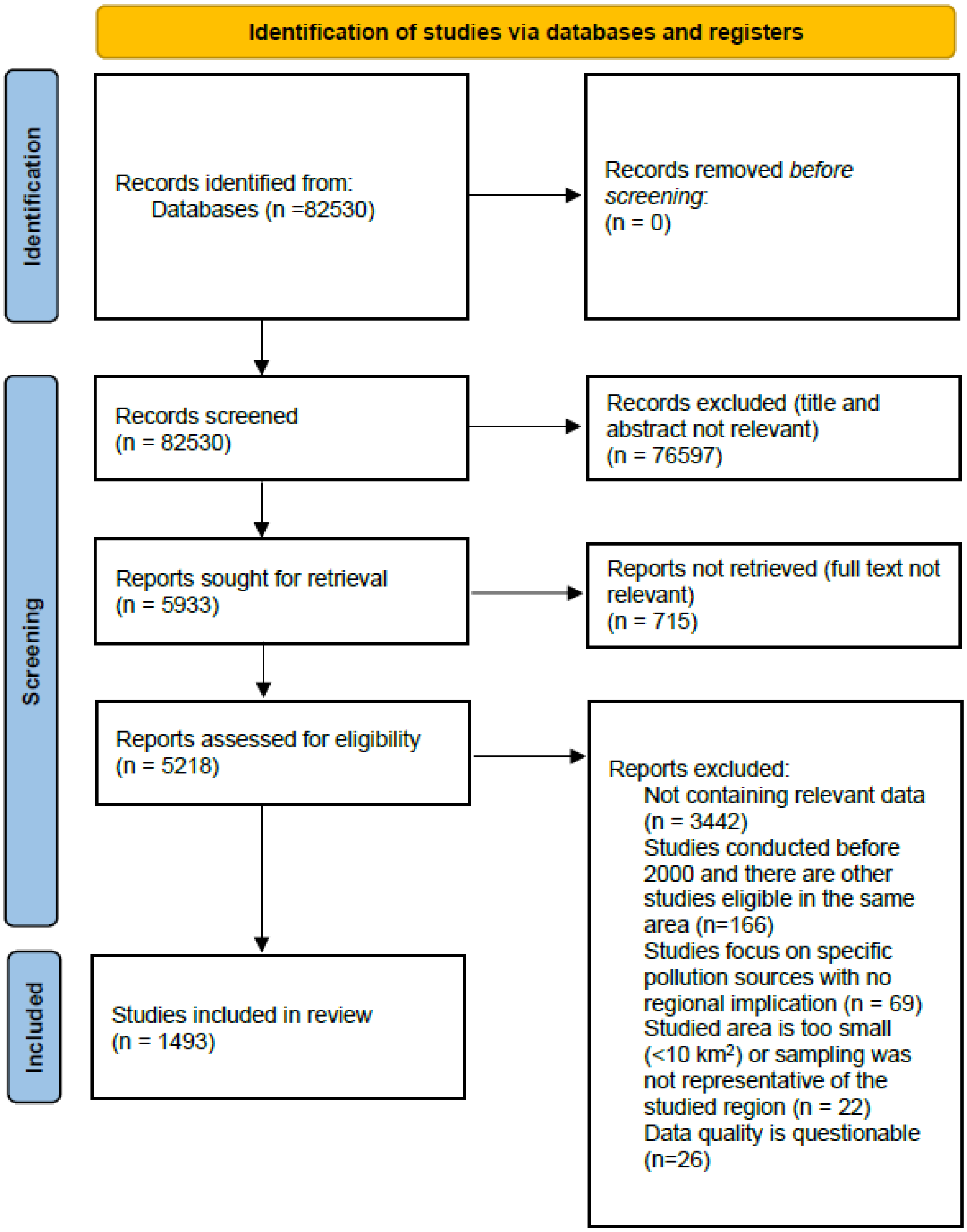

Fig. S1 PRISMA 2020 flow diagram for searches of database.

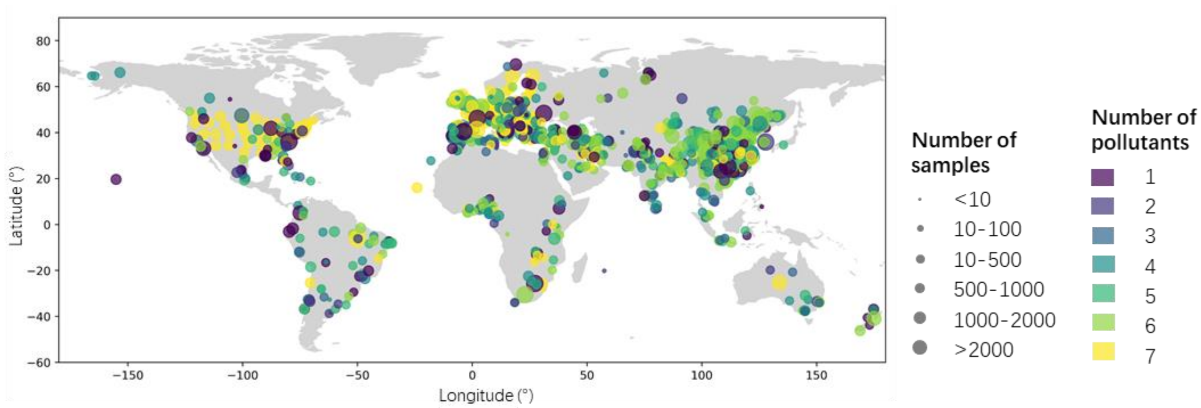

Fig. S2 Distribution of soil samples. Samples are relatively densely distributed in China, Europe, and the US, and more sparsely distributed in Central and Northern Asia, Africa, Australia, and Latin America.

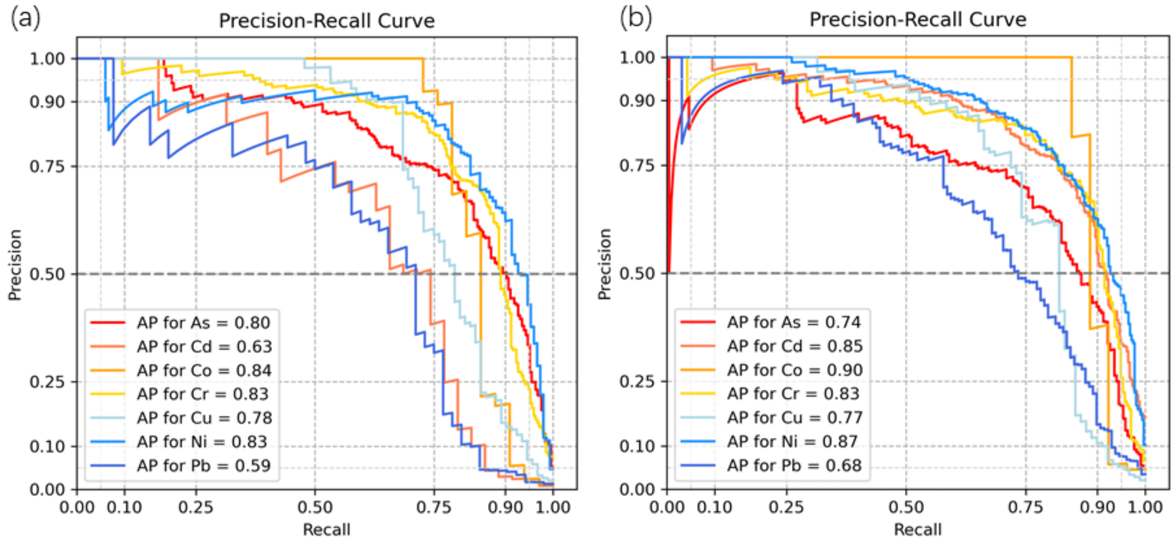

Fig. S3 Precision-recall curve and average precision for different metals. a) Models trained for human health and ecological threshold; b) Models trained for agricultural threshold.

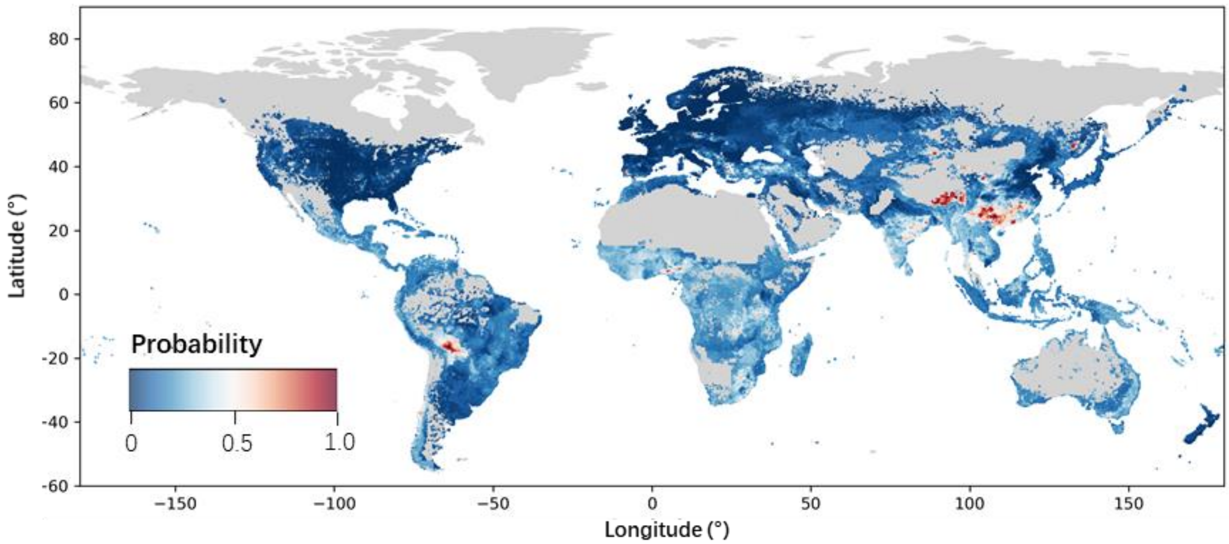

Fig. S4 Probability of As exceedance of agricultural threshold. Red showing high probability, and blue showing low probability. High probability is predicted for southwestern China, south and southeastern Asia, western Africa, and central parts of south America.

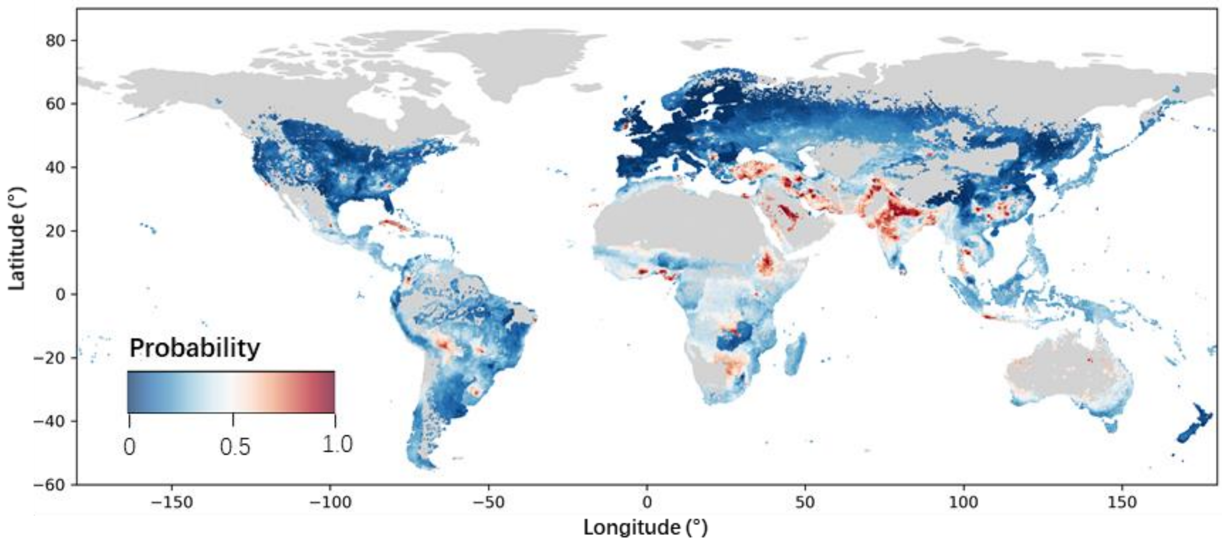

Fig. S5 Probability of Cd exceedance of agricultural threshold. Red showing high probability, and blue showing low probability. High probability is predicted for south Asia, the Middle-East, eastern Africa, and central America.

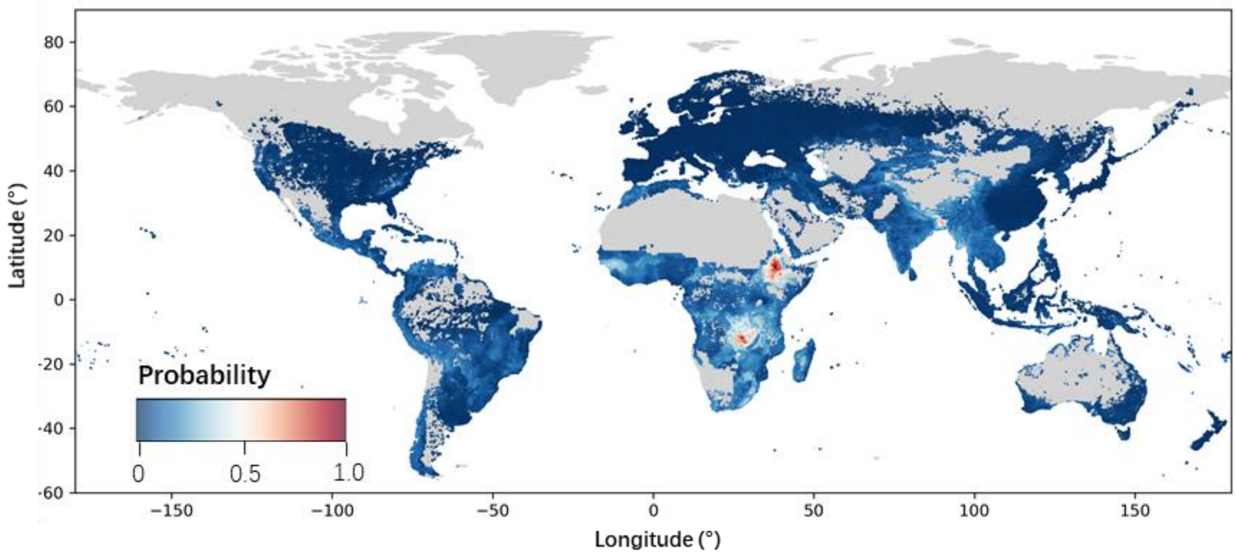

Fig. S6 Probability of Co exceedance of agricultural threshold. Red showing high probability, and blue showing low probability. High probability is predicted for eastern Africa.

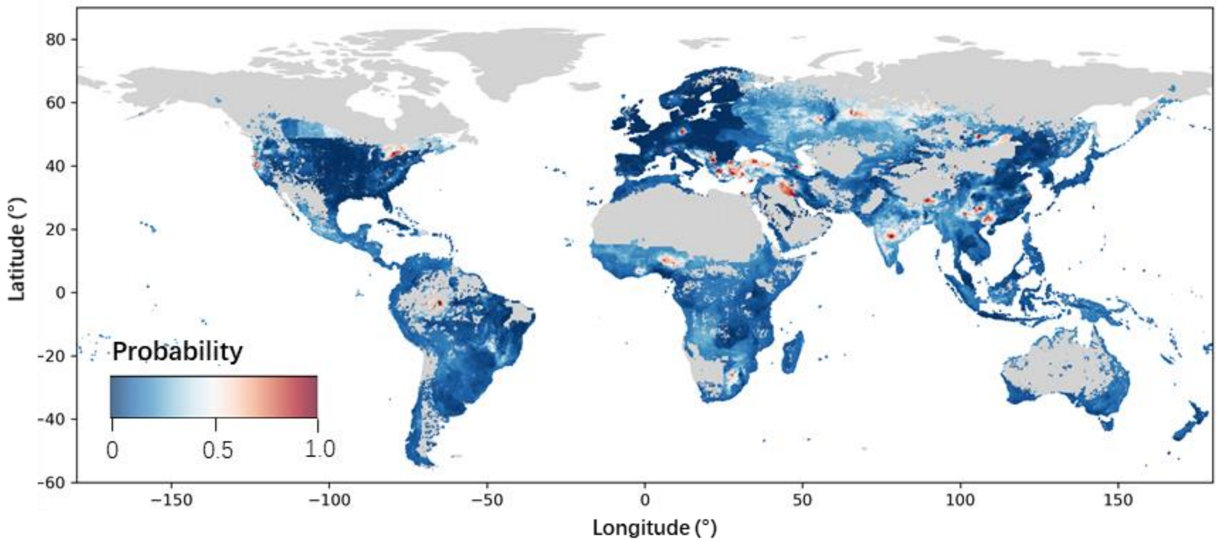

Fig. S7 Probability of Cr exceedance of agricultural threshold. Red showing high probability, and blue showing low probability. High probability is predicted for the Middle-East and subarctic Russia.

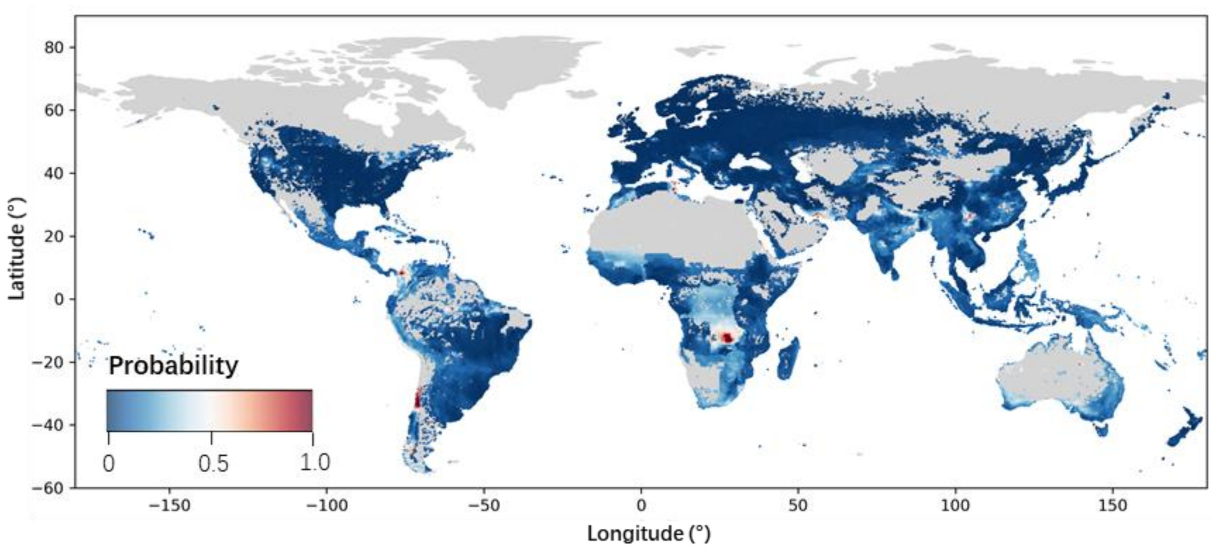

Fig. S8 Probability of Cu exceedance of agricultural threshold. Red showing high probability, and blue showing low probability. High probability is predicted for Zambia.

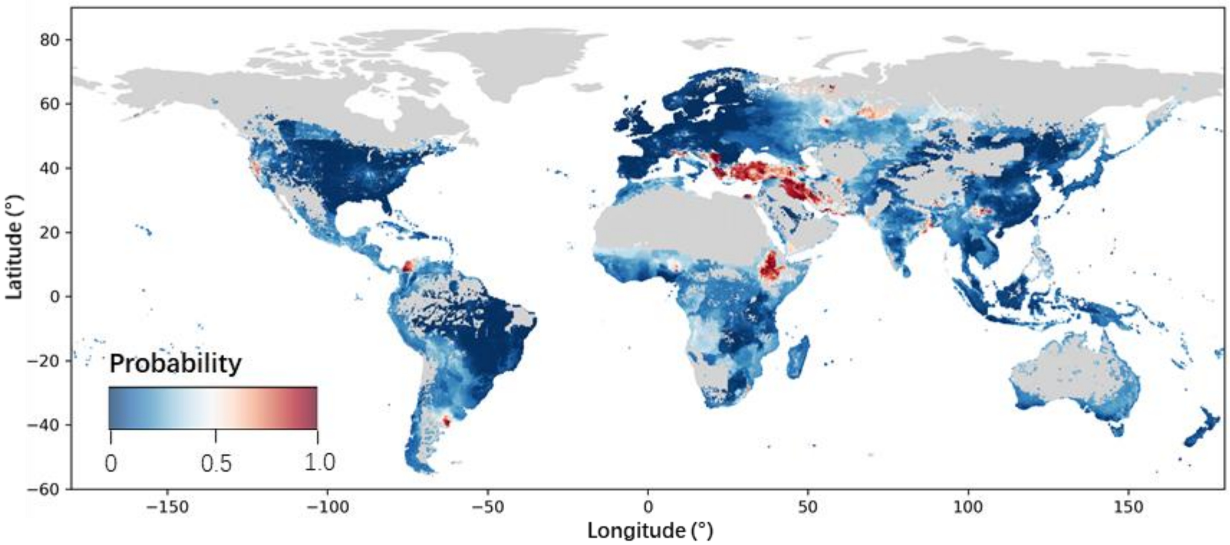

Fig. S9 Probability of Ni exceedance of agricultural threshold. Red showing high probability, and blue showing low probability. High probability is predicted for the Middle-East, eastern Africa, and Russia.

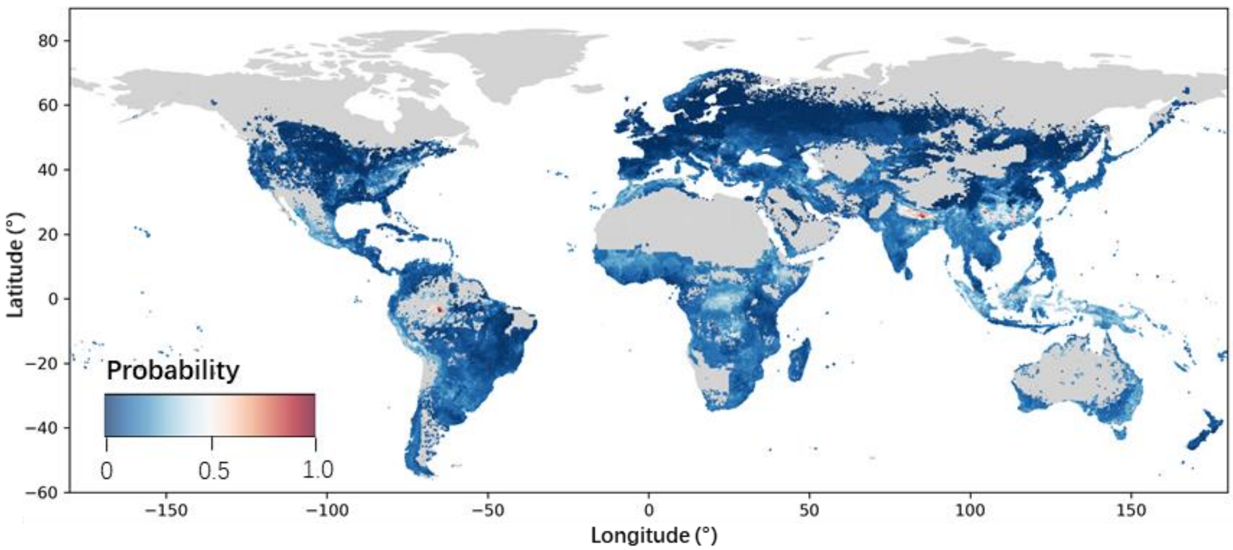

Fig. S10 Probability of Pb exceedance of agricultural threshold. Red showing high probability, and blue showing low probability. High probability is predicted for the Northern-India, and southern China.

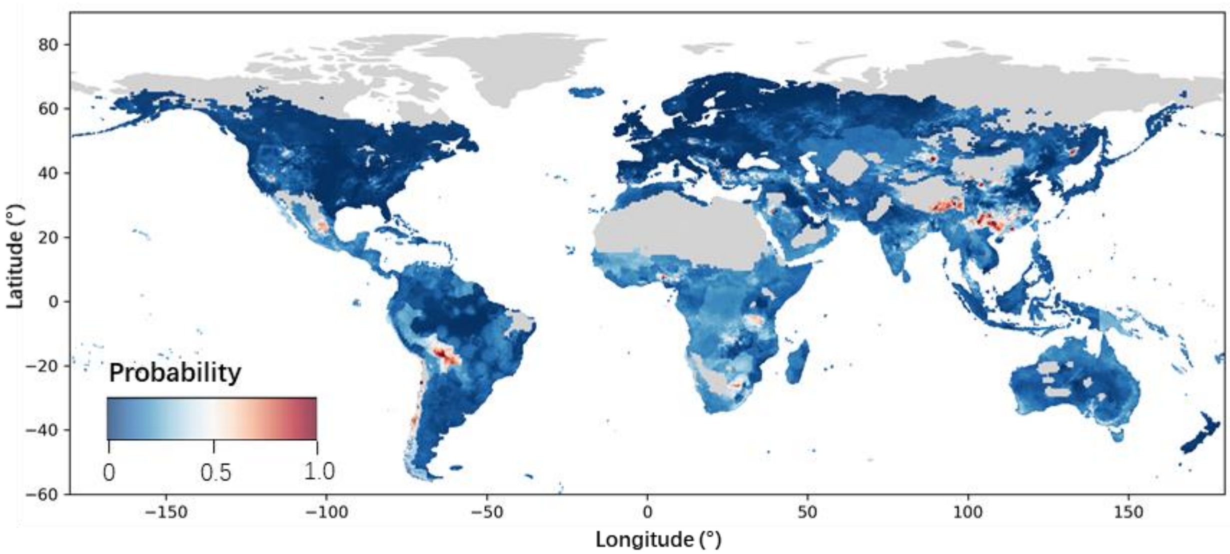

Fig. S11 Probability of As exceedance of human health and ecological threshold. Red showing high probability, and blue showing low probability. High probability is predicted for southwest China.

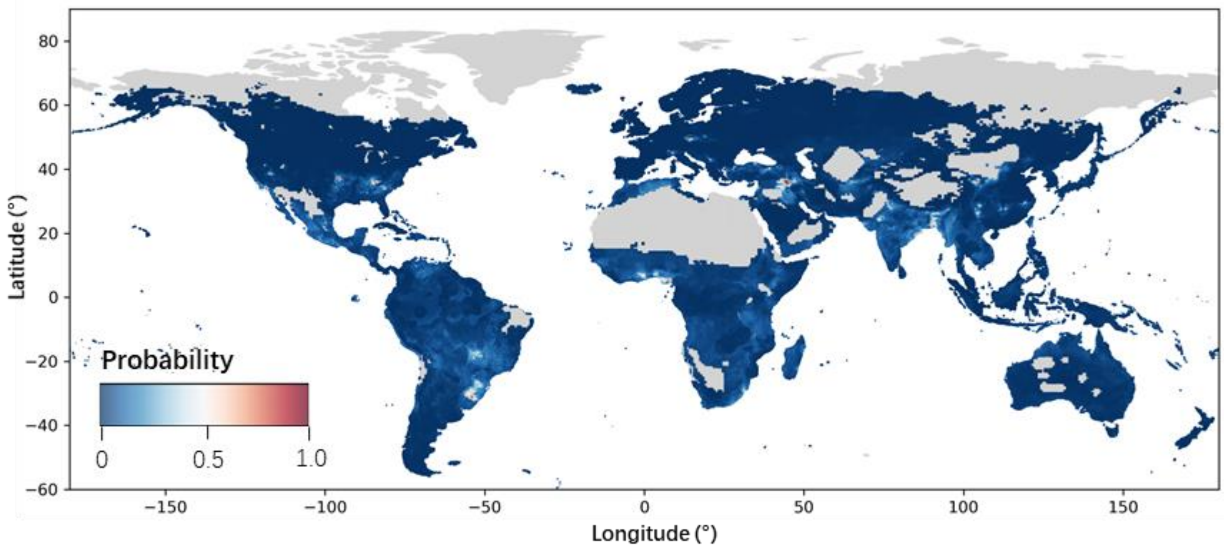

Fig. S12 Probability of Cd exceedance of human health and ecological threshold. Red showing high probability, and blue showing low probability. High probability is rarely predicted.

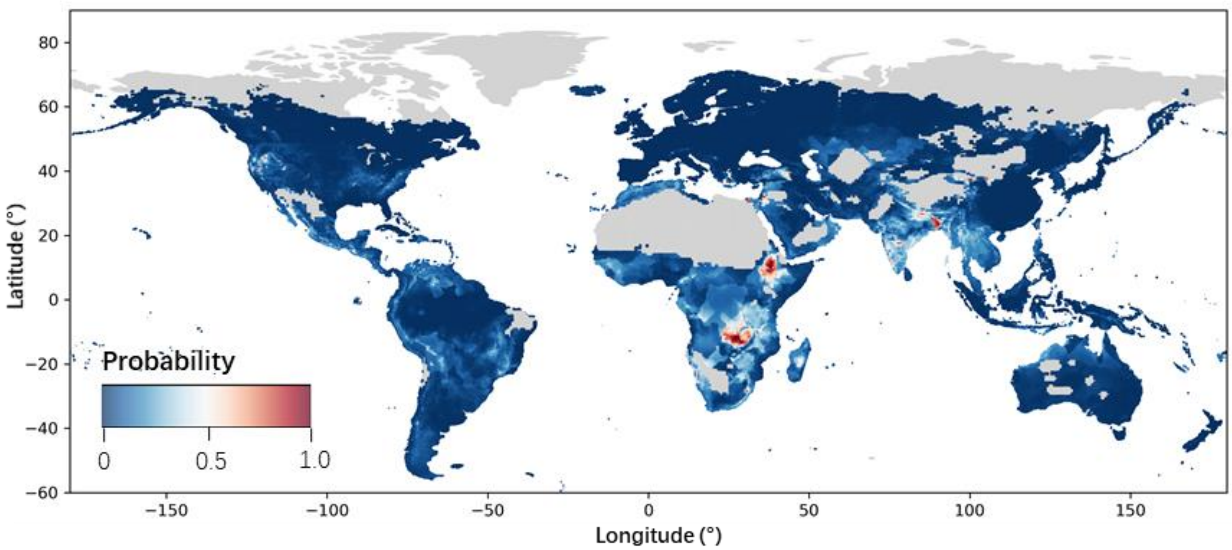

Fig. S13 Probability of Co exceedance of human health and ecological threshold. Red showing high probability, and blue showing low probability. High probability is predicted for south Asia, eastern Africa, and Zambia.

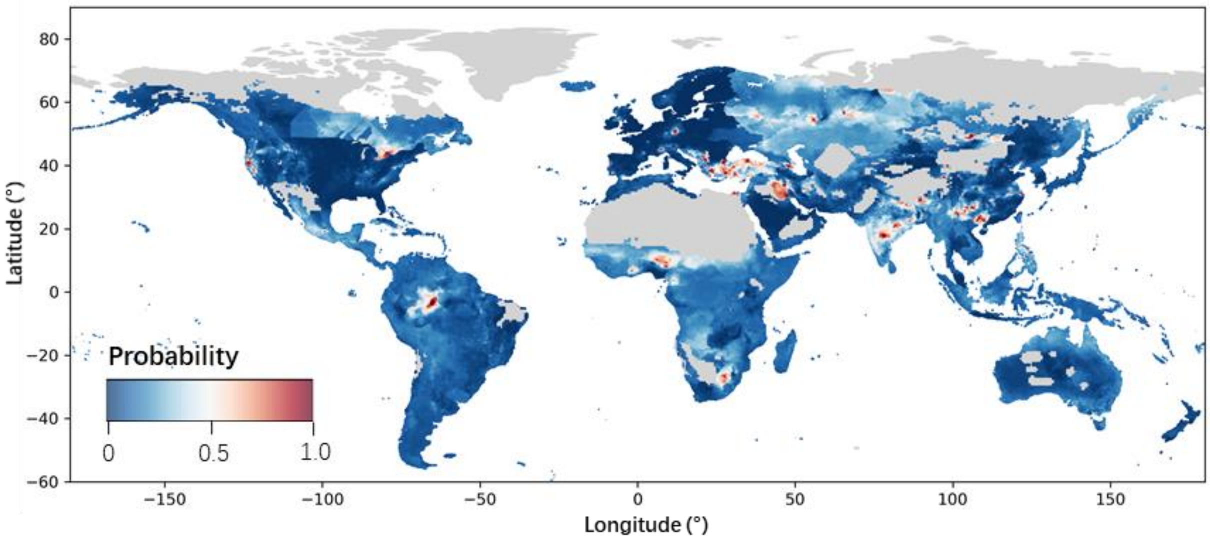

Fig. S14 Probability of Cr exceedance of human health and ecological threshold. Red showing high probability, and blue showing low probability. High probability is predicted for the Middle-East.

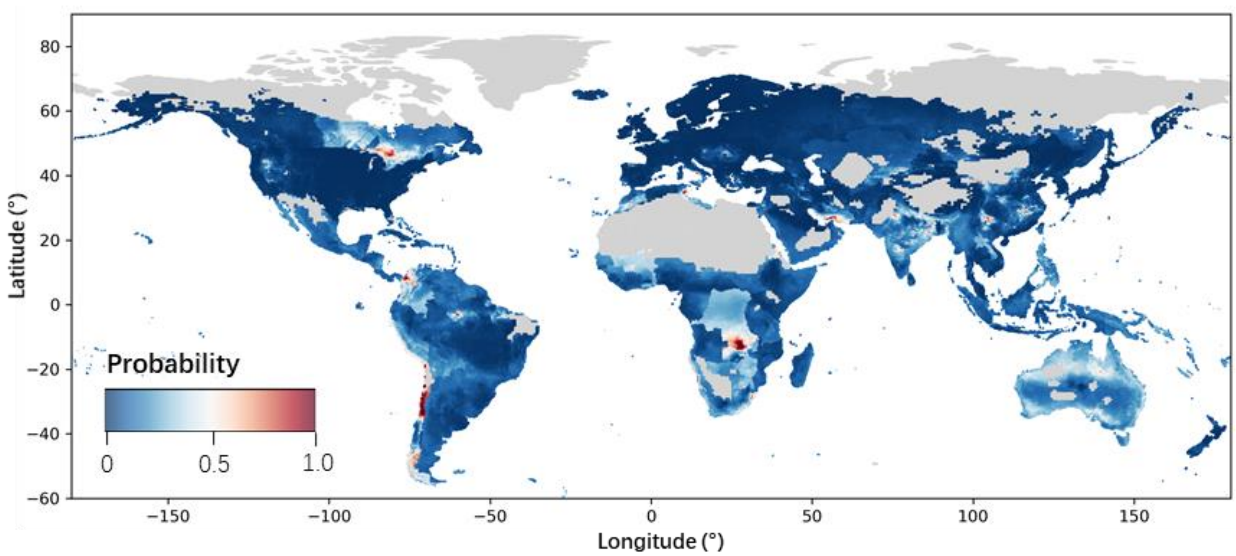

Fig. S15 Probability of Cu exceedance of human health and ecological threshold. Red showing high probability, and blue showing low probability. High probability is predicted for south Asia, Zambia, Chile, and central America.

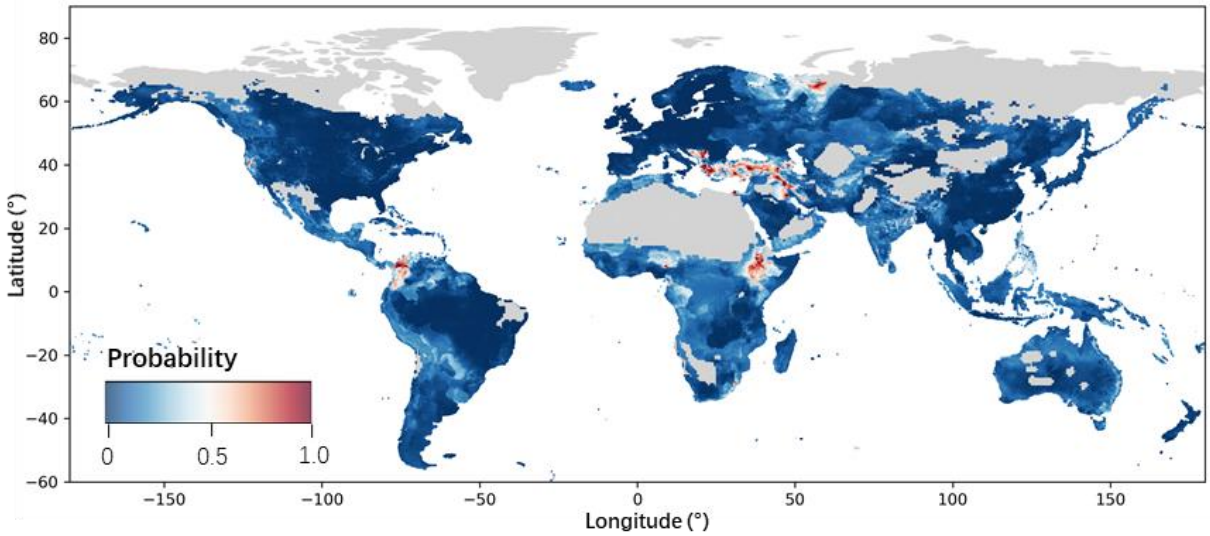

Fig. S16 Probability of Ni exceedance of human health and ecological threshold. Red showing high probability, and blue showing low probability. High probability is predicted for the Middle-East, and eastern Africa.

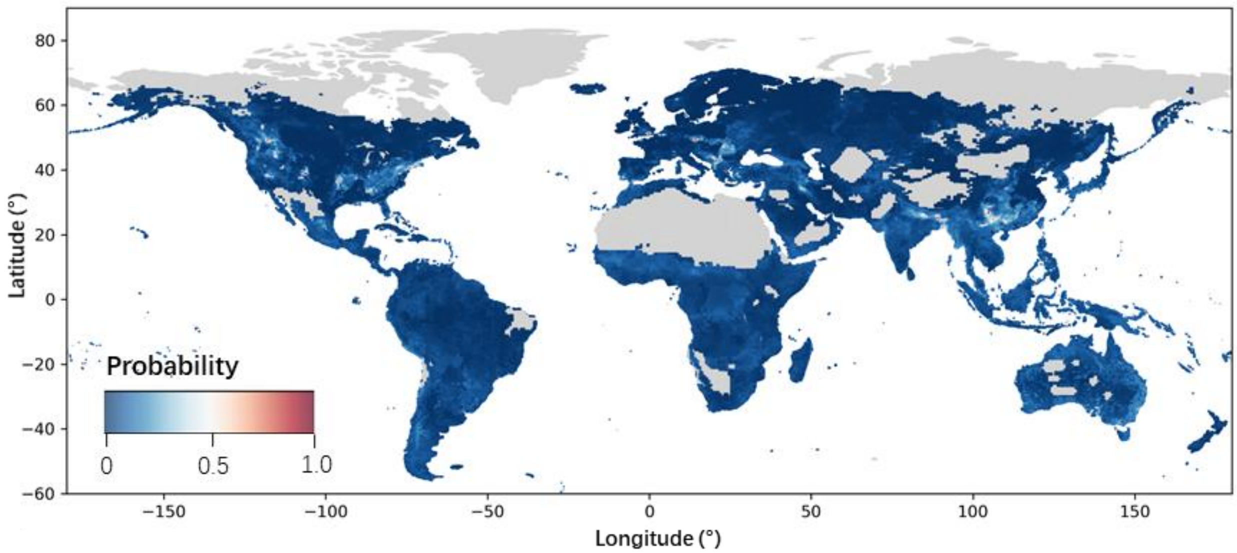

Fig. S17 Probability of Pb exceedance of human health and ecological threshold. Red showing high probability, and blue showing low probability.

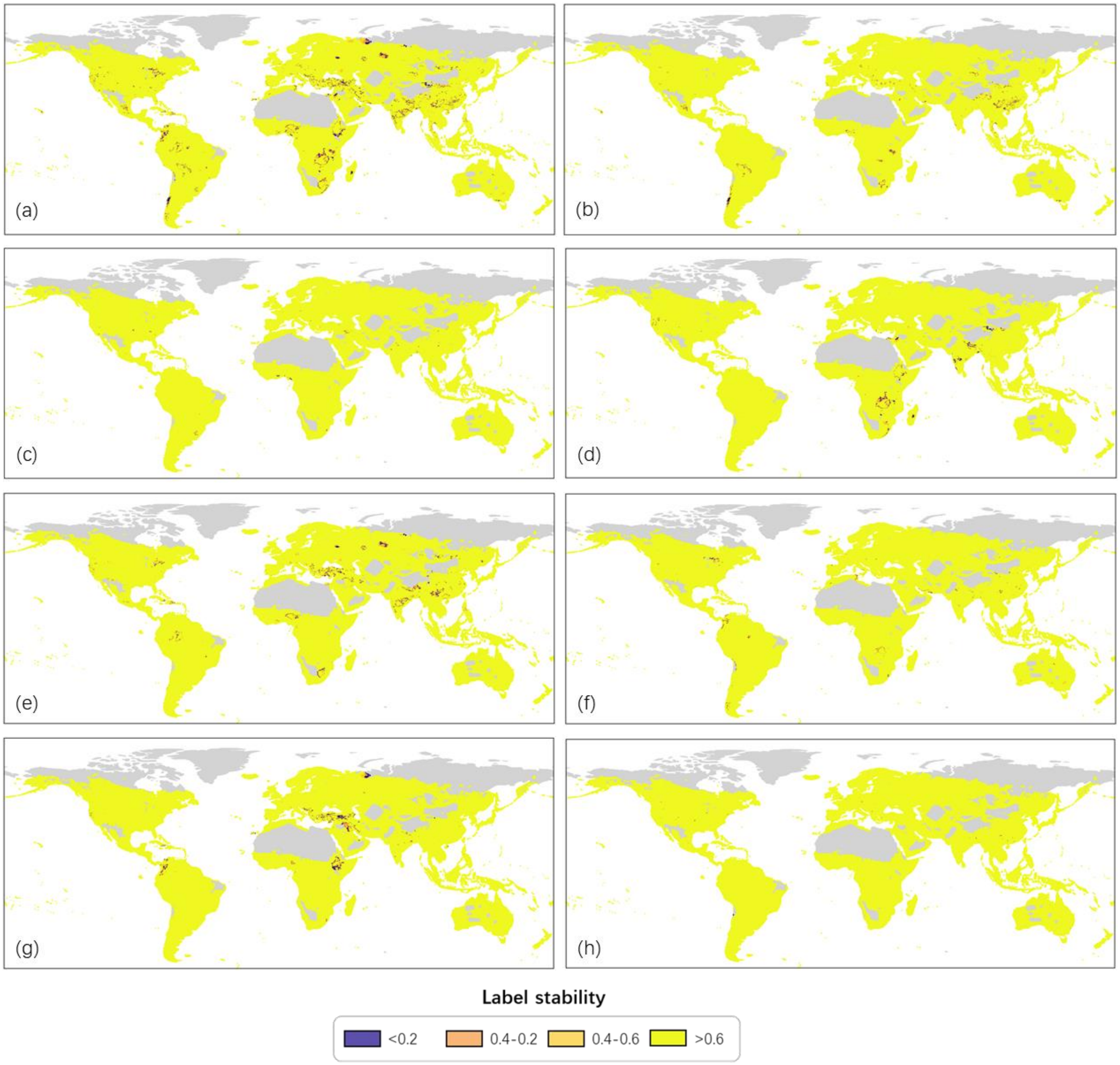

Fig. S18 Label stability for human health and ecological thresholds. (a) Total metals; (b) As; (c) Cd; (d) Co; (e) Cr; (f) Cu; (g) Ni; (h) Pb. High stability is observed for most of the areas, with some notable exceptions in discontinuous areas of northern Russia, south Asia, the Middle East, and eastern Africa; Grey=no data.

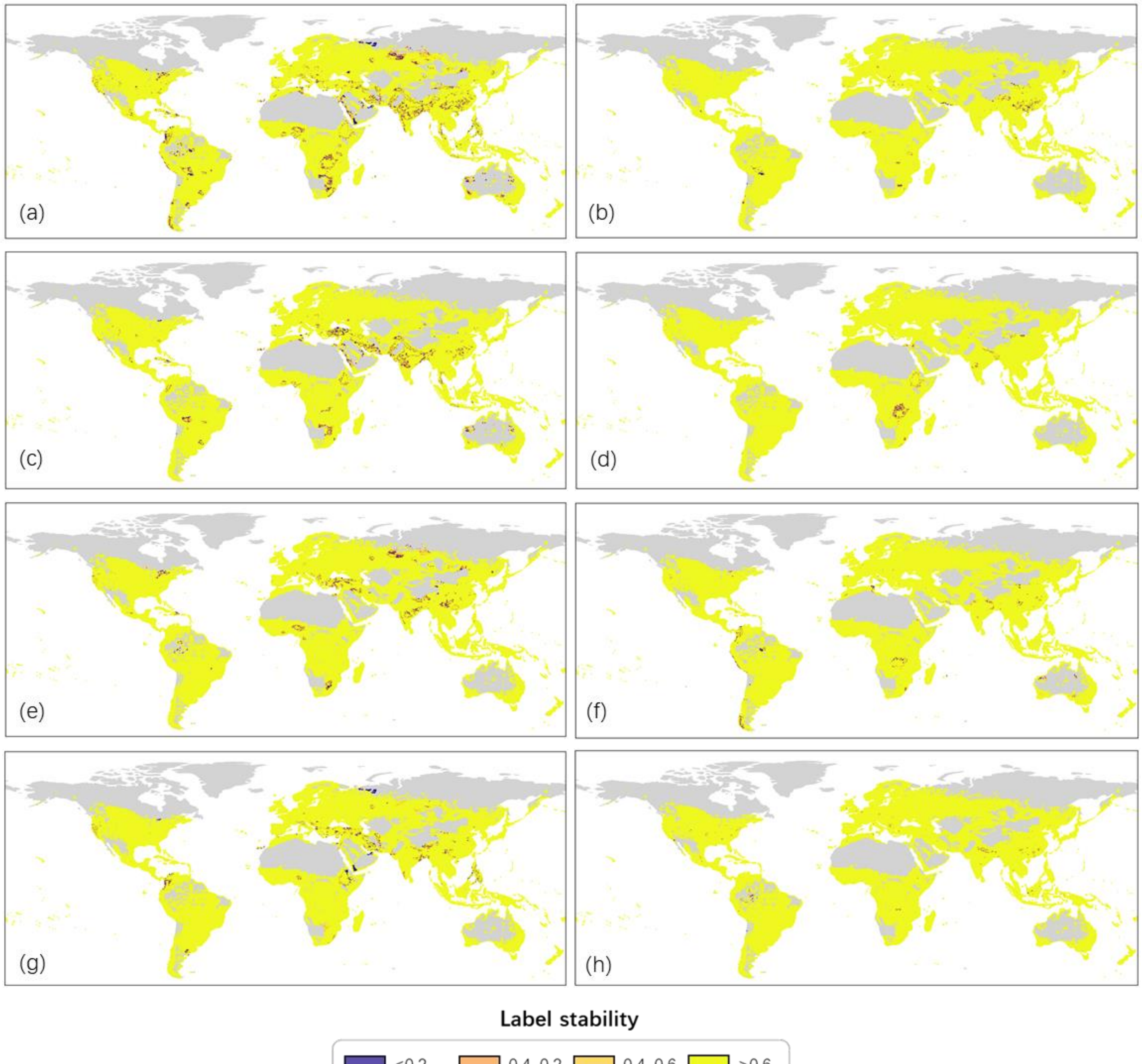

Fig. S19 Label stability for agricultural thresholds. (a) Total metals; (b) As; (c) Cd; (d) Co; (e) Cr; (f) Cu; (g) Ni; (h) Pb. High stability is observed for most of the areas, with most notable exceptions in northern Russia, but also discontinuous areas of east and south Asia, the Middle East, Africa, Latin America, and Australia; Grey=no data.

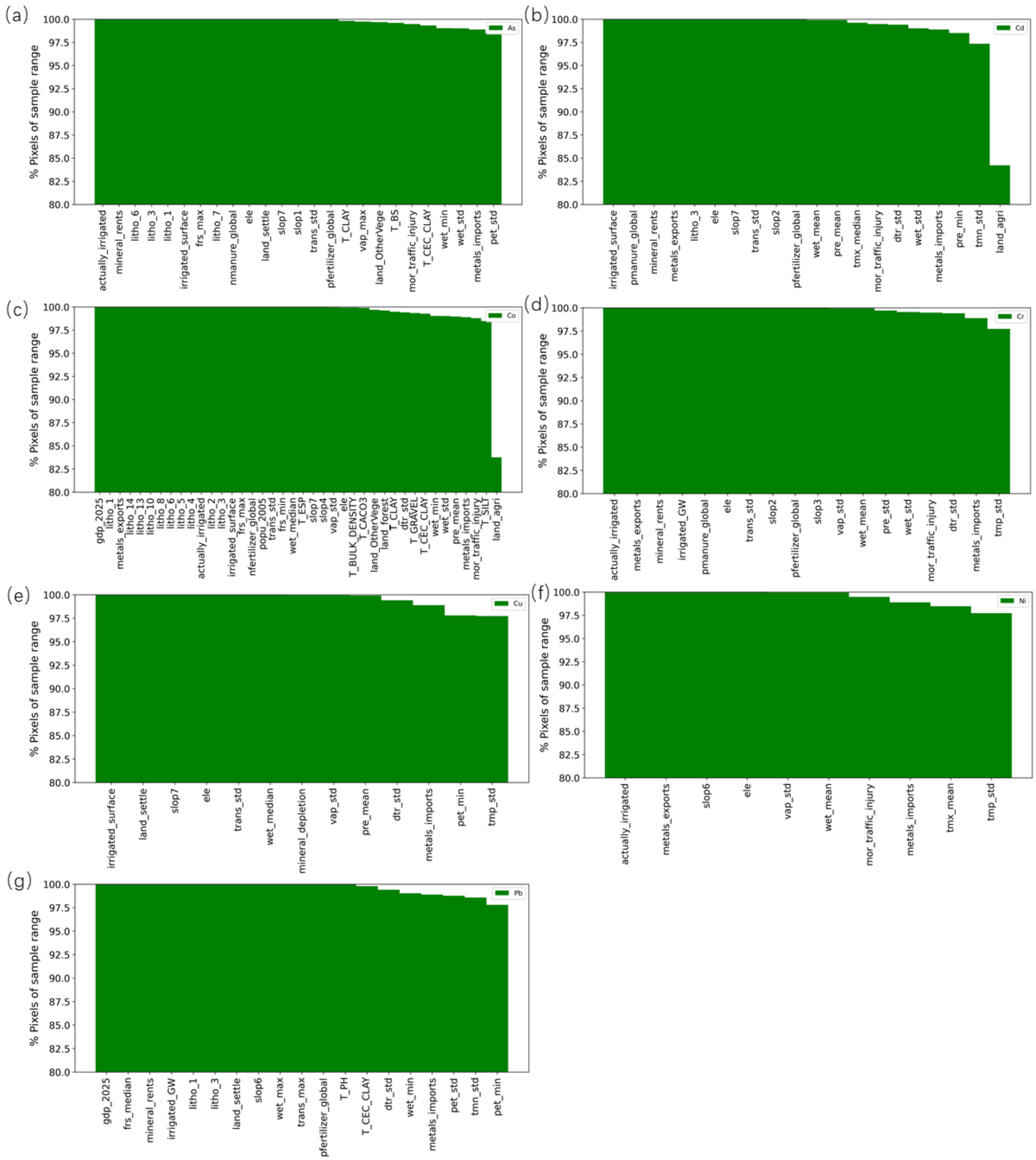

Fig. S20 Percentage of pixels interpolated for each variable. (a) As; (b) Cd; (c) Co; (d) Cr; (e) Cu; (f) Ni; (g) Pb. For most variables, the percentage is well above 95%, indicating good coverage.

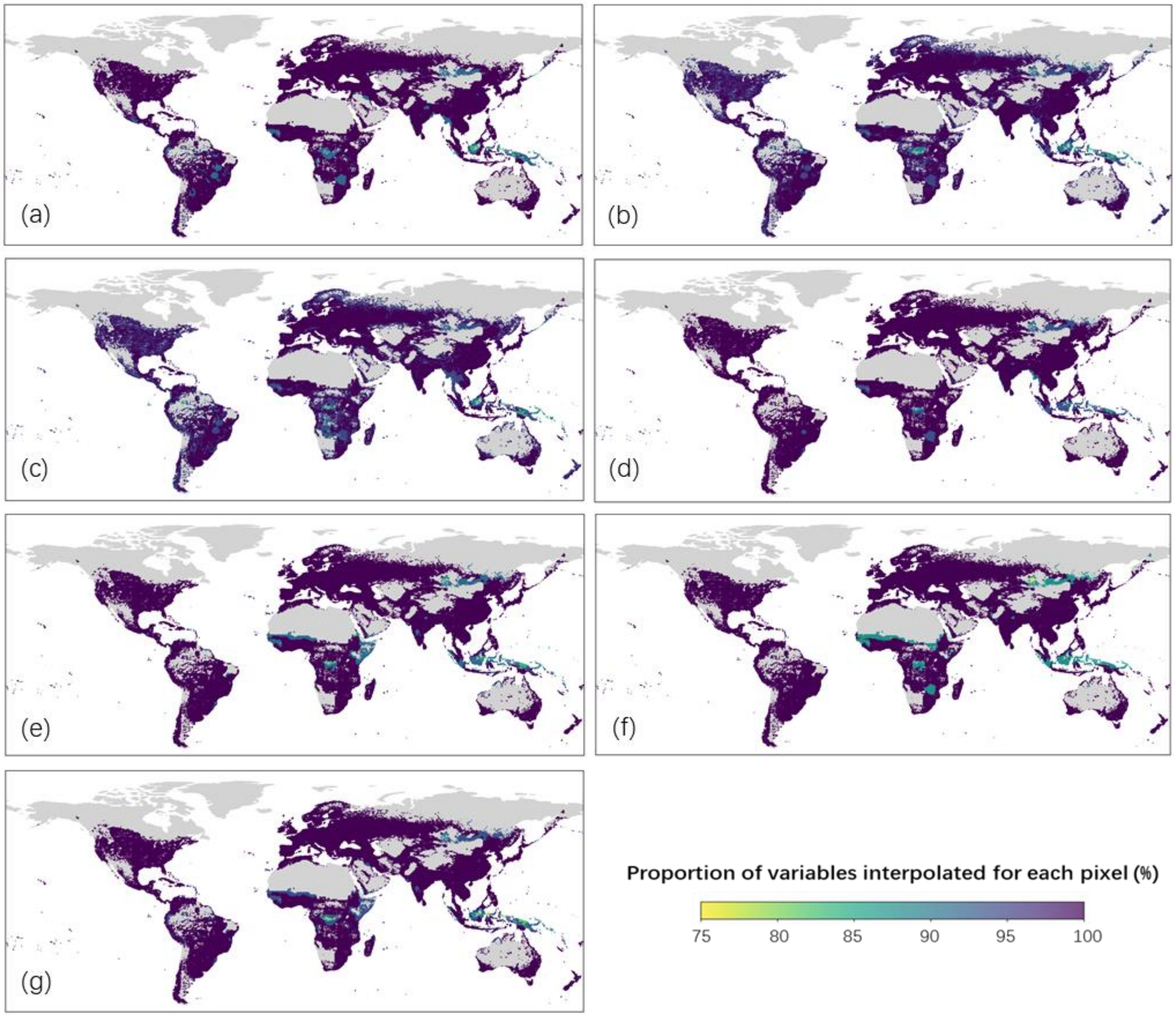

Fig. S21 Proportion of variables interpolated for each pixel. (a) As; (b) Cd; (c) Co; (d) Cr; (e) Cu; (f) Ni; (g) Pb. For most areas, the percentage is well above 95%, indicating good coverage.

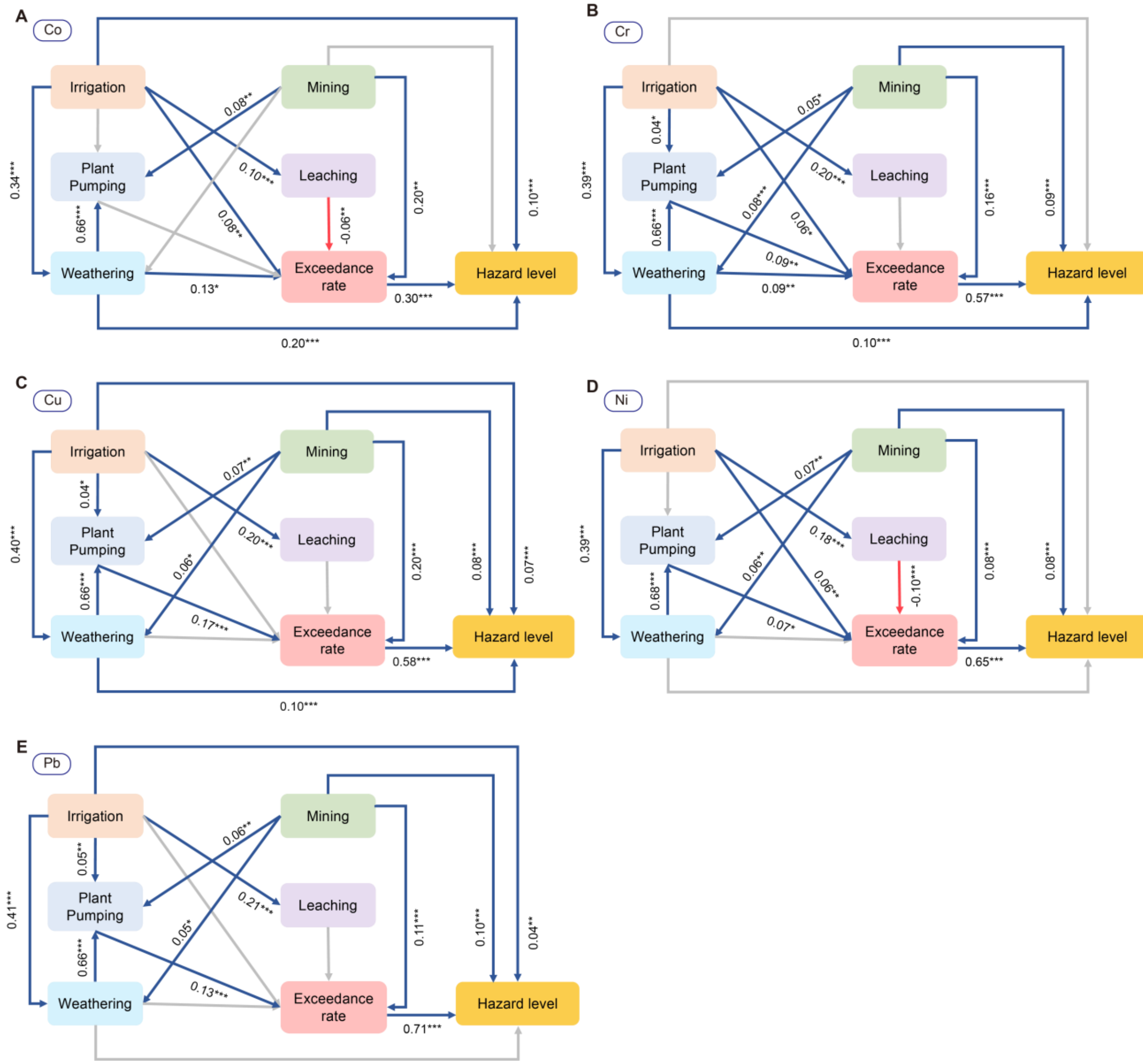

Fig. S22 Structure equation models for the other toxic metals. (a) Co; (b) Cr; (c) Cu; (d) Ni; (e) Pb. “***” denotes significant effect with p value less than 0.001; “**” denotes significant effect with p value less than 0.01; “*” denotes significant effect with p value less than 0.05, “.” denotes significant effect with p value less than 0.1.

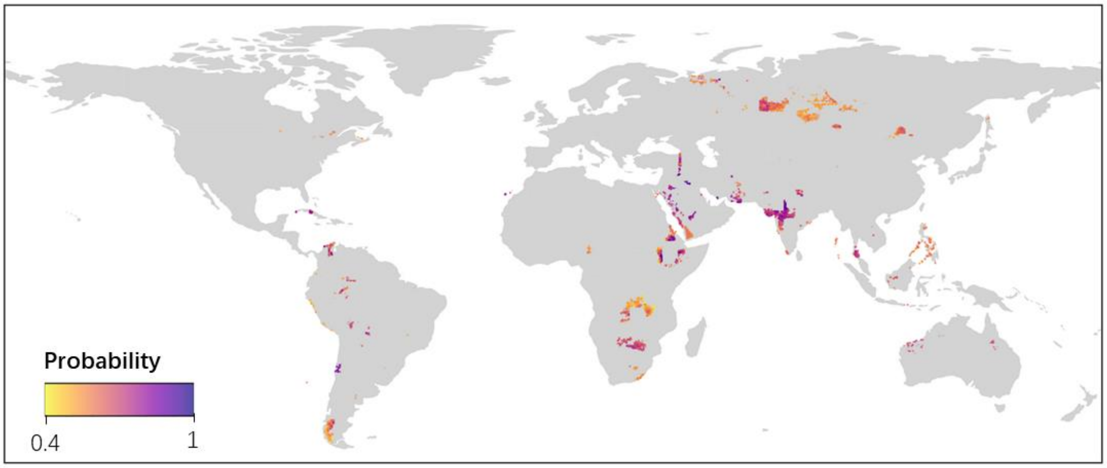

Fig. S23 Undocumented areas with potential exceedance of agricultural threshold predicted by machine learning models. Many of these areas are located in Africa, South Asia, Russia, and the Mid-East.

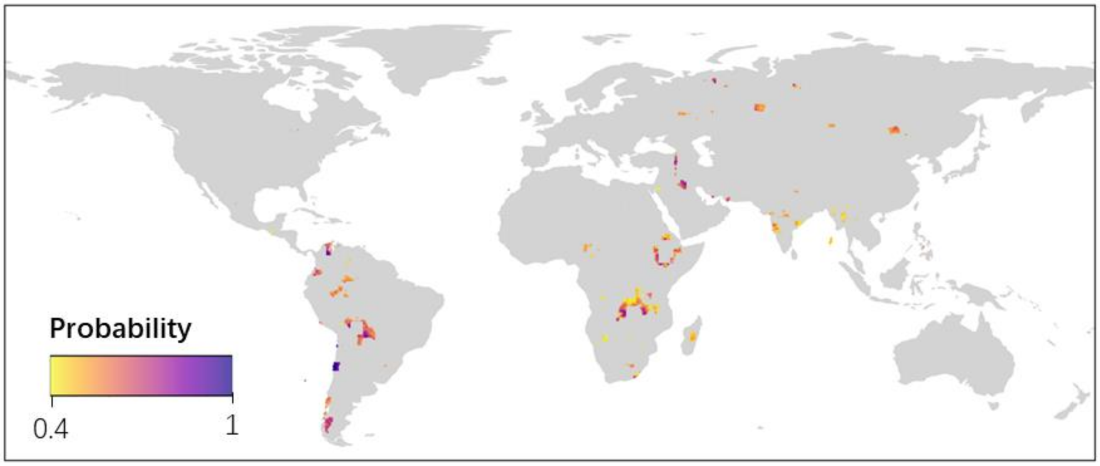

Fig. S24 Undocumented areas with potential exceedance of human health and ecological threshold predicted by machine learning models. Many of these areas are located in Africa and southern America.

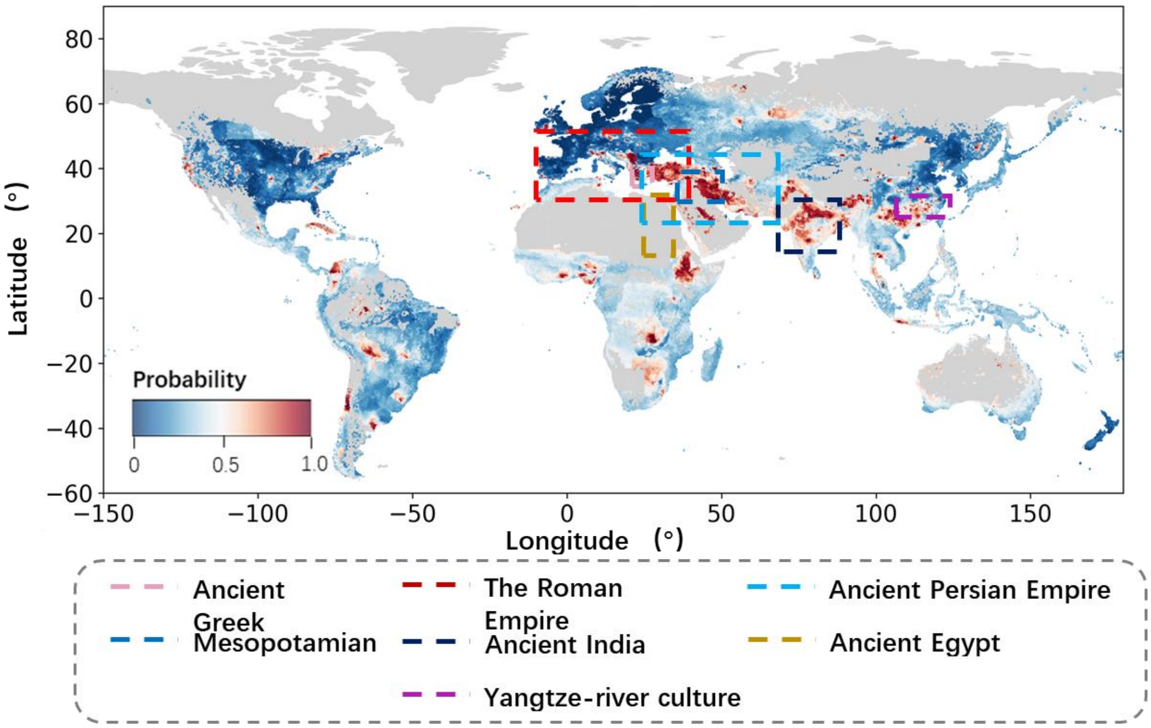

Fig. S25 Ancient cultures alone the metal enriched corridor. These cultures have largely overlapped with the metal enriched zone, and may have contributed to metal accumulation in history.

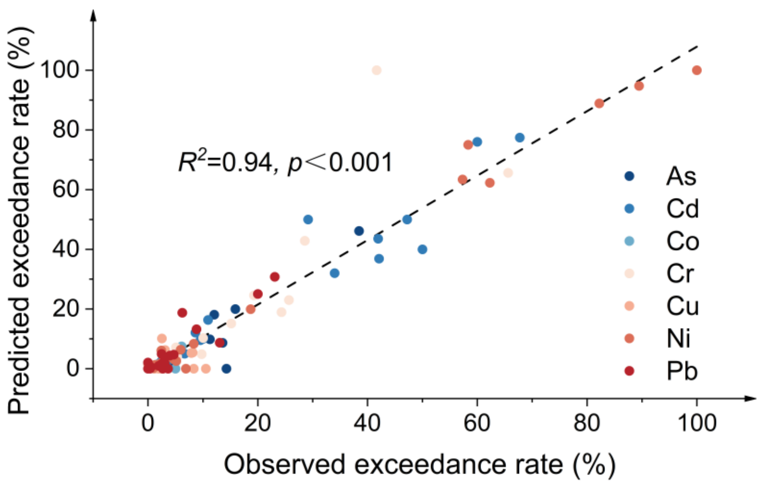

Fig. S26 Observed versus predicted metal exceedance rates.

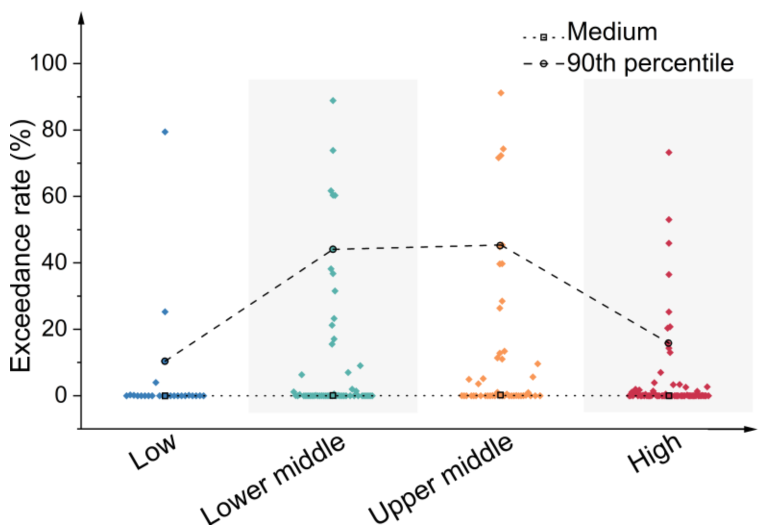

Fig. S27 Predicted exceedance rates in countries of various income levels.

4 Supplementary Tables

Table S1 Agricultural threshold (AT) and human health and ecological threshold (HHET) for soil pollution

Toxic metals	Agricultural threshold	human health and ecological threshold	Unit
As	20	20	mg/kg
Cd	1	6	mg/kg
Cr	100	100	mg/kg
Co	40	36.5	mg/kg
Cu	91	100	mg/kg
Ni	51	89	mg/kg
Pb	100	200	mg/kg

* the thresholds are derived from regulatory thresholds from 11 countries (see Table S2 and text S1.2)

Table S2 Regulatory thresholds from 11 countries (all units are in mg/kg)

Country	Threshold type	As	Cd	Cr	Co	Cu	Ni	Pb
Human health and ecological thresholds								
Austria	Trigger value-residential	20	2	50		100	70	100
Austria	Intervention value-residential	50	10	250		600	140	500
Belgium	Screening level-special	45	2	130		200	100	200
Belgium	Screening level-residential	110	6	300		400	470	700
Belgium	Screening level-industrial	300	30	800		800	700	2500
Belgium	Cleanup level-nature area	45	2	130		200	100	200
Belgium	Cleanup level-residential	110	6	300		400	470	700
Belgium	Cleanup level-recreational	200	15	500		500	550	1500
Belgium	Cleanup level-industrial	300	30	800		800	700	2500
Canada	SQG-residential/parkland	12	10	64	50	63	45	140
Canada	SQG-commercial	12	22	87	300	91	89	260
Canada	SQG-industrial	12	22	87	300	91	89	600
China	Intervention level -residential	120	47		190	8000	600	800
China	Intervention level -industrial	140	172		350	36000	2000	2500
China	Screening level -residential	20	20		20	2000	150	400
China	Screening level -industrial	60	65		70	18000	900	800
Denmark	Ecotoxicological soil quality criteria	10	0.3	50		30	10	50
Finland	Threshold value	5	1	100	20	100	50	60
Finland	Lower guideline value	50	10	200	100	150	100	200
Finland	Upper guideline value	100	20	300	250	200	150	750
France	VDSS	19	10	65	120	95	70	200
France	VDI-usage sensible	37	20	130	240	190	140	400
France	VDI-usage non sensible	120	60	7000	1200	950	900	2000
Germany	Triggering level-Playing grounds	25	10	200			70	200
Germany	Triggering level-residential	50	20	400			140	400
Germany	Triggering level-Park	125	50	1000			350	1000
Germany	Triggering level-industrial	140	60	1000			900	2000
Italy	Limit values-residential	20	2	150	20	120	120	100
Italy	Limit values-industrial	50	15	800	250	600	500	1000

Country	Threshold type	As	Cd	Cr	Co	Cu	Ni	Pb
Netherland	Target value	29	0.8	100	9	36	35	85
Netherland	Intervention value	55	12	380	240	190	210	530
US	US-RSL-residential	0.68	71	120000	23	3100	1500	400
US	US-RSL-industrial	3	980	1800000	350	47000	22000	800
	25 percentile	20	6	100	36.5	100	89	200
Agricultural thresholds								
Austria	Trigger value-agricultural	20	1	100		100	60	100
Belgium	Clean-up level-agricultural	45	2	130		200	100	200
Canada	SQG-agricultural	12	1.4	64	40	63	45	70
China	Screening level-agricultural *	30	0.45	237.5		131.25	52.5	126.25
China	Intervention level-agricultural *	143	2.625	987.5				650
	25 percentile	20	1	100	40	91	51	100

* represents average for paddy field and non-paddy field under various pH ranges

Table S3 Parameter tuning range

Parameters	**Range**
n_estimators	[1000, 2000,3000]
max_depth	[10,15,20]
criterion	["gini","entropy"]
min_samples_split	[2, 5, 10]
min_samples_leaf	[1, 3, 8]

Table S4 Optimal hyperparameter settings for different metals

Pollutant	criterion	Estimators	Max depth	Min Samples leaf	Min samples split
Human health and ecological thresholds					
As	entropy	2000	20	1	2
Cd	gini	3000	20	1	2
Co	entropy	2000	20	1	2
Cr	entropy	1000	20	1	2
Cu	entropy	1000	20	1	5
Ni	entropy	2000	20	1	2
Pb	entropy	3000	20	1	10
Agricultural thresholds					
As	entropy	3000	20	1	5
Cd	entropy	1000	20	1	5
Co	entropy	2000	15	1	2
Cr	entropy	3000	20	1	2
Cu	entropy	1000	20	1	2
Ni	entropy	1000	20	1	2
Pb	entropy	2000	20	1	5

Table S5 Probability cut-offs to determine whether grids were affected by toxic metals

Toxic metal	Human health and ecological thresholds	Agricultural thresholds
As	0.55	0.66
Cd	0.54	0.64
Co	0.42	0.41
Cr	0.51	0.51
Cu	0.63	0.5
Ni	0.57	0.54
Pb	0.7	0.59

Table S6 Model performance of ERT for all toxic metals

	As	Cd	Co	Cr	Cu	Ni	Pb
Human health and ecological thresholds							
BA	0.87	0.81	0.88	0.87	0.85	0.89	0.81
F1-score	0.75	0.66	0.82	0.79	0.71	0.79	0.62
Sensitivity	0.76	0.63	0.76	0.75	0.71	0.79	0.62
Specificity	0.99	1.00	1.00	0.99	1.00	1.00	1.00
AUC	0.87	0.81	0.88	0.87	0.85	0.89	0.81
KIA	0.74	0.65	0.82	0.78	0.71	0.79	0.61
AP	0.80	0.63	0.84	0.83	0.78	0.83	0.59
Agricultural thresholds							
BA	0.86	0.91	0.92	0.88	0.85	0.89	0.80
F1-score	0.72	0.78	0.86	0.79	0.74	0.79	0.64
Sensitivity	0.74	0.83	0.85	0.77	0.71	0.80	0.62
Specificity	0.99	0.98	1.00	0.99	1.00	0.99	0.99
AUC	0.86	0.91	0.92	0.88	0.85	0.89	0.80
KIA	0.71	0.76	0.86	0.78	0.74	0.78	0.63
AP	0.74	0.85	0.90	0.83	0.77	0.87	0.68

BA represents balanced accuracy. AUC is the area under the curve of receiver operating characteristic. KIA means Cohen's kappa coefficient and AP denotes average precision.

Table S7 Soil toxic metal exceedance in European Union and China

Region	Study	Estimation Method	Threshold Value type	Total	As	Cd	Co	Cr	Cu	Ni	Pb
European Union	this study	Machine learning	Health risk Threshold		20	6	36.5	100	100	89	200
			Exceedance rate	2.9%	0.8%	0.1%	0.1%	1.2%	0.2%	1.5%	0.2%
			Agricultural threshold		20	1	40	100	91	51	100
			Exceedance rate-AT	5.1%	0.6%	0.7%	0.0%	1.0%	0.4%	3.3%	0.6%
	Toth, 2016a (*100*)	Simple proportion of sample	Threshold value		5	1	20	100	100	50	60
			Exceedance rate-Agriculture	**58.1%** [1]							
			Exceedance rate-All	**53.3%** [1]		5.5%	4.5%	2.7%			
			Lower Guidance value		50	10	100	200	150	100	200
			Exceedance ratel	**6.2%** [1]	0.8%		0.38%	1.1%			
			Higher Guidance value		100	20	250	300	200	150	750
			Exceedance rate-Agriculture	**2.6%** [1]							
			Exceedance rate-All	**2.4%** [1]							
	Toth, 2016b (*11*)	Regression kriging	Threshold value		5	1	20	100	100	50	60
			Exceedance rate	28.3% [2]	25.5%	0.3%	1.0%	0.5%	0%	3.9%	0.2%
			Health risk threshold [3]		20	6	36.5	100	100	89	200
			Exceedance rate	1.4% [2]	0.1%	0.0%	0.1%	0.5%	0.0%	1.1%	0.0%
			Agricultural threshold [3]		20	1	40	100	91	51	100
			Exceedance rate	4.2% [2]	0.1%	0.3%	0.0%	0.5%	0.0%	3.8%	0.0%
China	this study	Machine learning	Health risk Threshold		20	6	36.5	100	100	89	200
			Exceedance rate	13.8%	11.0%	0.2%	0.8%	5.2%	0.6%	0.2%	0.2%

Region	Study	Estimation Method	Threshold Value type	Total	As	Cd	Co	Cr	Cu	Ni	Pb
			Agricultural threshold		20	1	40	100	91	51	100
			Exceedance rate	12.8%	6.8%	4.2%	0.1%	4.1%	1.0%	2.1%	1.0%
	Chen, 2015 (*32*)	Simple proportion of sample	Grade 1 threshold		15	0.2		90	35	40	35
			Exceedance rate	**66.8%** [4]	16.9%	27.7%		14.7%	15.8%	13.6%	20.0%
			Grade 2 threshold		30	0.6		200	200	50	300
			Exceedance rate	**9.6%** [4]	4.0%	3.8%		1.3%	0.3%	6.1%	0.2%
	MEP, 2014 (*10*)	Simple proportion of sample	Soil quality standard [5]		30	0.3		250	150	50	300
			Exceedance rate	**11.8%** [6]	2.7%	7.0%		1.1%	2.1%	4.8%	1.5%

[1] This exceedance rate also includes exceedances of mercury, zinc, and vanadium
[2] This exceedance rate was derived using average toxic metal concentrations extracted from the TIF files provided by the study
[3] Thresholds used in the present study
[4] Combined exceedance was derived by adding individual toxic metal exceedance and multiply a factor derived from MEP, 2014
[5] The mean of standards for various pH range and soil type is listed.
[6] Combined exceedance was derived by subtracting exceedance rates of mercury, zinc, and organic pollutants from the overall exceedance rate

Table S8 Difference among three exceedance inference methods

Potentially toxic elements	Human health and ecological thresholds			Agricultural threshold		
	Simple proportion of sample exceedance	Inference with aggregated probability[1]	Inference with average concentration	Simple proportion of sample exceedance	Inference with aggregated probability[1]	Inference with average concentration
As	12.1%	4.8%	7.7%	12.1%	4.8%	7.7%
Cd	0.8%	0.5%	0.9%	8.8%	7.2%	10.2%
Co	3.8%	1.2%	1.4%	3.2%	1.2%	1.3%
Cr	12.8%	5.7%	7.6%	12.8%	5.7%	7.6%
Cu	3.2%	2.0%	2.5%	3.8%	2.1%	2.8%
Ni	3.2%	3.2%	3.9%	10.1%	7.3%	9.0%
Pb	2.7%	0.7%	1.2%	7.3%	2.1%	3.3%

[1]This exceedance rate is calculated based on Beta distribution and Bayesian inference mentioned in Section 1.3.

Table S9 Parameters for health assessment of toxic metals through ingestion, inhalation and dermal pathways

Parameters	Discription	Unit	Value
IngR	Ingestion rate	mg/day	100
InhR	Inhalation rate	m^3/day	20
EF	Exposure frequency	Days/year	350
ED	Exposure duration	Years	30
BW	Body weight	kg	70
AT	Average timing	Days	10950
SA	Skin area	cm^2	5700
ABS	Dermal adsorption factor	No unit	0.03 (As) 0.001 (other metal)
AF	Adherence factor of soil	mg/cm^3/day	0.07
PEF	Particulate emission factor	m^3/kg	1.36×10^{9}
CF	units conversion factor	kg/mg	1×10^{-6}

Source: (*126*, *127*)

Table S10 Reference doses for different pathways

Pollutants	Ingestion	Inhalation	Dermal contact
As	3.00×10^{-4}	1.23×10^{-4}	3.01×10^{-4}
Cd	1.00×10^{-3}	1.00×10^{-3}	1.00×10^{-5}
Co	3.00×10^{-4}	6.00×10^{-6}	-
Cr	3.00×10^{-3}	2.86×10^{-5}	5.00×10^{-5}
Cu	4.00×10^{-2}	-	1.20×10^{-2}
Ni	2.00×10^{-2}	2.06×10^{-2}	5.40×10^{-3}
Pb	3.50×10^{-3}	3.52×10^{-3}	5.25×10^{-4}

Source: (*128*)