nature portfolio

Peer Review File

Global and regional patterns of soil metal(loid) mobility and associated risks

Corresponding Author: Professor Zhang Lin

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The study highlighted the dynamic nature of metal(loid) mobility, provided insights for mitigating global soil pollution and mostly demonstrated the potential of Machine Learning in transforming soil metal(loid) fractionation analysis.

Excellent work and nice presentation. The tables and figures are self-explanatory.
However, some minor things are suggested to be addressed.

1. L26, L28, L30, L32, L77, L245, L359, L384, L427, L434,………. and throughout both manuscript and Supplementary files, please replace the personal pronoun with i.e., in this study/this investigation/ or make the sentences passive where appropriate.
2. L 65: Although advanced analytical 'techniques', such as …. Here only one technique is mentioned, would you mind checking if there are any typos or correcting it?
3. L70: "of these experiments", here which experiments are referred to? Please clarify it.

4. L291, the word 'Should' can be replaced by 'might'

5. In L401 and 402, how can you say, "indicating its high generalization capability"? Is it considering the R2 values? If so, can you please put a reference to support this statement? Or if anything else, please elaborate on that.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The manuscript reports the development of a machine learning (ML) model to predict the potential for metal(loid) mobility globally, with outcomes presented in terms of both fraction of metal(loid)s likely to be mobile and, where metal(loid) concentrations are available, the absolute mobile concentrations expected. The scope and scale of the work is considerable, for which the authors are to be commended. The work is also presented and communicated well. I am not aware of anything similar having been attempted previously. In this review I have focused primarily on the purpose, applicability and value of the model developed – I have only very limited experience of ML and allied disciplines and so am not able to review much of the technical methodology with any confidence. I have the following questions and comments:

1. Perhaps my most significant question regards the potential application of the model. Why is it necessary to obtain comprehensive and fine-resolution data at continental and global scales (Lines 62-64), and what purpose would this have? I can see that there may be value at a very high level in discussions about land use and planning, and there is speculation on this in lines 434-438 (albeit somewhat vague), but at the site scale this would likely only act as an indicator at best. I would not expect it to provide sufficient information or confidence to support local level or site-specific management or remediation as it is highly likely there will be substantial variation in data within the 5x5km resolution (e.g. variability in local geology, soil amendments, impact of site development) that will affect soil properties including carbon levels and pH. However, with site-

specific data (including metal(loid) concentrations) this could have value in acting as an alternative to carrying out sequential extractions. The purpose and value of the work should be clarified considerably.

2. It is somewhat disingenuous to say that the study has identified significant risks, when this study doesn't appear to consider the actual risk to human health or other factors. The assessment of the potential mobility of contaminants is part of the risk analysis, but not the whole. The mobility may be significant if contamination were to occur, at levels sufficient to cause harm and if receptors (assuming a typical source-pathway-receptor risk analysis approach) are present. The analysis comes up with very high proportions (e.g. 82%) of land at risk from metal(loid) mobility, but this doesn't mean the metal(loids) are actually there and that a significant risk would arise even if they were. I think some statements are potentially misleading – for example, the caption for Figure 2 states that this shows "Risk hotspots of prevalent metal(loid)s" – this is not the case, as the mobility of contaminants is only part of the risk analysis. I'm of the opinion that there is a high level of value in the work done, though the exact nature of this value needs to be clarified accurately throughout.

3. The data used for training and validation of the model isn't uniform globally or within countries. There is quite a good level of coverage, but there are gaps (to be expected). However, are the training data sufficiently representative of global conditions that the resulting model is accurate everywhere? The model is validated on a subset of the data used, but given the validation datasets in particular appear to be limited to particular geographical regions (Supplementary Table 8 suggests a large proportion is from China, whilst all the lab tests are from China) then is this valid for other parts of the world?

4. Spatial disparities in EU soil (lines 207-243): In Figure 4, comparing the total Pb content in France and the UK there looks (visually at least) to be significant difference in content from the map (Fig 4a) whilst Fig 4b suggests there is only a small difference in mean content. The majority of the UK, for example, appears to be at the highest level of Pb (coloured orange, appears >100 mg/kg?) whilst France is much more variable. Depending on where the UK RGVs lie (see next comment), this may also imply that the vast majority of the UK is above typical RGVs in EU countries – is this really true? Does this result from the colour scale of lead content used being apparently non-uniform, with 40-50% of the Pb range considered being coloured orange? Unless there is good reason for this (e.g. if it relates to the RGVs in these regions somehow), perhaps the colour spectrum used should vary more linearly to more clearly distinguish the levels of contamination.
a. Line 216 - it is stated that the majority of RGVs for Pb are 50 or 100 mg/kg, yet in Supplementary Table 2 the baseline RGV is 150 mg/kg for the EU. Can this be explained and/or corrected?
b. Is there any external validation of the total and/or mobile metal(loid)s contents found in this EU-focused analysis? For example, how does independent literature stemming from the EU compare to the model outputs?

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
This manuscript presents a novel application of machine learning (ML) for modeling the global mobility of soil metal(loid)s. From the perspective of a soil scientist, this study marks the first attempt to investigate soil metal(loid)s mobility on a global scale, leading to several intriguing findings. The work is of significant interest to a broad audience and is suitable for publication in Nature Communications. However, I propose the following suggestions for improvement before its acceptance:
1. The authors should elaborate on the rationale for selecting the BCR sequential extraction method. Does this method adequately represent metal(loid)s mobility? Please provide further discussion.
2. What exactly does metal(loid)s mobility entail? How is this mobility connected to the actual risk of metal(loid)s pollution? The authors should clarify this linkage.
3. The global application of the trained ML model is predicated on the assumption of consistent total metal(loid)s content. While this is a widely accepted approach in the absence of global distribution data, it raises an important question: How might changes in total metal(loid)s content impact the global distribution, particularly the identification of hotspots? This warrants further investigation.
4. The uncertainty analysis should be conducted for the ML modeling.
5. The reliability of the ML model for global applications should be addressed. Can the model consistently provide stable predictions across diverse regions?
6. Why was the EU chosen for this study? Large-scale metal(loid)s data are also available for other countries, such as China. The authors should justify this choice.
7. In Lines 419–421, the authors mention selecting 22 metal(loid)s with their RGVs within the data range. Could the authors explain why this selection criterion was applied in the context of the study?

(Remarks on code availability)

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
All the raised comments have been addressed. Based on the other reviewers' comments it can be accepted for publication.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
The authors have done well to comprehensively address my questions and concerns with clarity. In specific response to certain rebuttals:
• Point 1: My question on the potential application of the model has been well addressed by the new addition in the discussion. I am of the opinion that this could be slightly enhanced in the abstract and introduction, but this is only a minor recommendation and I am, overall, happy with the response.
• Point 2: The redefinition of risk as mobilisation risk has worked well and seems reasonable. This has been addressed consistently throughout the manuscript and the revised text is much clearer on this point.
• Point 3: The clarification on ML training and validation seems to be reasonable. As this isn't my speciality, then if the other reviewers are happy with this then this will be acceptable to me.

Minor comment:
• Line 275 – "exiting" should be 'existing'.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
I have carefully read the author's responses to the reviewers and the revised manuscript. The authors have madea satisfactory revision on the manuscript and addressed all my comments and suggestions. I am satisfied with the responses and revisions, so I suggest accepting the manuscript for publication in NC.

(Remarks on code availability)

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

The study highlighted the dynamic nature of metal(loid) mobility, provided insights for mitigating global soil pollution and mostly demonstrated the potential of Machine Learning in transforming soil metal(loid) fractionation analysis.

Excellent work and nice presentation. The tables and figures are self-explanatory.

However, some minor things are suggested to be addressed.

1. L26, L28, L30, L32, L77, L245, L359, L384, L427, L434,………. and throughout both manuscript and Supplementary files, please replace the personal pronoun with i.e., in this study/this investigation/ or make the sentences passive where appropriate.

2. L 65: Although advanced analytical 'techniques', such as …. Here only one technique is mentioned, would you mind checking if there are any typos or correcting it?

3. L70: "of these experiments", here which experiments are referred to? Please clarify it.

4. L291, the word 'Should' can be replaced by 'might'

5. In L401 and 402, how can you say, "indicating its high generalization capability"? Is it considering the R2 values? If so, can you please put a reference to support this statement? Or if anything else, please elaborate on that.

Reviewer #2 (Remarks to the Author):

The manuscript reports the development of a machine learning (ML) model to predict the potential for metal(loid) mobility globally, with outcomes presented in terms of both fraction of metal(loid)s likely to be mobile and, where metal(loid) concentrations are available, the absolute mobile concentrations expected. The scope and scale of the work is considerable, for which the authors are to be commended. The work is also presented and communicated well. I am not aware of anything similar having been attempted previously. In this review I have focused primarily on the purpose, applicability and value of the model developed - I have only very limited experience of ML and allied disciplines and so am not able to review much of the technical methodology with any confidence. I have the following questions and comments:

1. Perhaps my most significant question regards the potential application of the model. Why is it necessary to obtain comprehensive and fine-resolution data at continental and global scales (Lines 62-64), and what purpose would this have? I can see that there may be value at a very high level in discussions about land use and planning, and there is speculation on this in lines 434-438 (albeit somewhat vague), but at the site scale this would likely only act as an indicator at best. I would not expect it to provide sufficient information or confidence to support local level or site-specific management or remediation as it is highly likely there will be substantial variation in data within the 5x5km resolution (e.g. variability in local geology, soil amendments, impact of site development) that will affect soil properties including carbon levels and pH. However, with site-specific data (including metal(loid) concentrations) this could have value in acting as an alternative to carrying out sequential extractions. The purpose and value of the work should be clarified considerably.

2. It is somewhat disingenuous to say that the study has identified significant risks, when this study doesn't appear to consider the actual risk to human health or other factors. The assessment of the potential mobility of contaminants is part of the risk analysis, but not the whole. The mobility may be significant if contamination were to occur, at levels sufficient to cause harm and if receptors (assuming a typical source-pathway-receptor risk analysis approach) are present. The analysis comes up with very high proportions (e.g. 82%) of land at risk from metal(loid) mobility, but this doesn't mean the metal(loids) are actually there and that a significant risk would arise even if they were. I think some statements are potentially misleading - for example, the caption for Figure 2 states that this shows "Risk hotspots of prevalent metal(loid)s" - this is not the case, as the mobility of contaminants is only part of the risk analysis. I'm of the opinion that there is a high level of value in the work done, though the exact nature of this value needs to be clarified accurately

throughout.

3. The data used for training and validation of the model isn’t uniform globally or within countries. There is quite a good level of coverage, but there are gaps (to be expected). However, are the training data sufficiently representative of global conditions that the resulting model is accurate everywhere? The model is validated on a subset of the data used, but given the validation datasets in particular appear to be limited to particular geographical regions (Supplementary Table 8 suggests a large proportion is from China, whilst all the lab tests are from China) then is this valid for other parts of the world?

4. Spatial disparities in EU soil (lines 207-243): In Figure 4, comparing the total Pb content in France and the UK there looks (visually at least) to be significant difference in content from the map (Fig 4a) whilst Fig 4b suggests there is only a small difference in mean content. The majority of the UK, for example, appears to be at the highest level of Pb (coloured orange, appears >100 mg/kg?) whilst France is much more variable. Depending on where the UK RGVs lie (see next comment), this may also imply that the vast majority of the UK is above typical RGVs in EU countries - is this really true? Does this result from the colour scale of lead content used being apparently non-uniform, with 40-50% of the Pb range considered being coloured orange? Unless there is good reason for this (e.g. if it relates to the RGVs in these regions somehow), perhaps the colour spectrum used should vary more linearly to more clearly distinguish the levels of contamination.

a. Line 216 - it is stated that the majority of RGVs for Pb are 50 or 100 mg/kg, yet in Supplementary Table 2 the baseline RGV is 150 mg/kg for the EU. Can this be explained and/or corrected?

b. Is there any external validation of the total and/or mobile metal(loid)s contents found in this EU-focused analysis? For example, how does independent literature stemming from the EU compare to the model outputs?

Reviewer #3 (Remarks to the Author):

This manuscript presents a novel application of machine learning (ML) for modeling the global mobility of soil metal(loid)s. From the perspective of a soil scientist, this study marks the first attempt to investigate soil metal(loid)s mobility on a global scale, leading to several intriguing findings. The work is of significant interest to a broad audience and is suitable for publication in Nature Communications. However, I propose the following suggestions for improvement before its acceptance:

1. The authors should elaborate on the rationale for selecting the BCR sequential extraction method. Does this method adequately represent metal(loid)s mobility? Please provide further discussion.

2. What exactly does metal(loid)s mobility entail? How is this mobility connected to the actual risk of metal(loid)s pollution? The authors should clarify this linkage.

3. The global application of the trained ML model is predicated on the assumption of consistent total metal(loid)s content. While this is a widely accepted approach in the absence of global distribution data, it raises an important question: How might changes in total metal(loid)s content impact the global distribution, particularly the identification of hotspots? This warrants further investigation.

4. The uncertainty analysis should be conducted for the ML modeling.

5. The reliability of the ML model for global applications should be addressed. Can the model consistently provide stable predictions across diverse regions?

6. Why was the EU chosen for this study? Large-scale metal(loid)s data are also available for other countries, such as China. The authors should justify this choice.

7. In Lines 419–421, the authors mention selecting 22 metal(loid)s with their RGVs within the data range. Could the authors explain why this selection criterion was applied in the context of the study?

Response to Reviewer #1

(Manuscript Number: NCOMMS-24-75613)

Overview:

The study highlighted the dynamic nature of metal(loid) mobility, provided insights for mitigating global soil pollution and mostly demonstrated the potential of Machine Learning in transforming soil metal(loid) fractionation analysis.

Excellent work and nice presentation. The tables and figures are self-explanatory.

However, some minor things are suggested to be addressed.

Response:

We thank the reviewer for the positive appreciation of our manuscript, particularly on its novelty and significance as also highlighted by the other two Reviewers. Revision based on your comments has been made to further improve the quality of this paper. Below are point-to-point responses in which we explicitly refer to the relevant lines in the revised manuscript/supplementary information. The comments appear in black, our responses appear in blue, and the texts quoted from the revised manuscript/supplementary information appear in red.

Specific Comments:

1. L26, L28, L30, L32, L77, L245, L359, L384, L427, L434,………. and throughout both manuscript and Supplementary files, please replace the personal pronoun with i.e., in this study/this investigation/ or make the sentences passive where appropriate.

Response:

We thank the reviewer for this comment. After a comprehensive review, we prefer to retain the active expressions in the revised manuscript, as this is a general writing style at Nature Communications. However, we do change the active expressions to passive ones in the Results section.

2. L65: Although advanced analytical 'techniques', such as …. Here only one technique is mentioned, would you mind checking if there are any typos or correcting it?

Response:

We agree with the suggestion here. We have changed the sentences with new supporting references.

Revised Statement:

"Although advanced analytical techniques, such as high-energy resolution X-ray absorption spectroscopy and isotope dilution, can accurately determine metal(loid) speciation[15,16], their specialized nature and limited availability restrict widespread application." (Page 3, line 65-68)

3. L70: “of these experiments”, here which experiments are referred to? Please clarify it.

Response:

Suggestion taken. We have changed the sentences to avoid vague statement.

Revised Statement:

“This is attributed to the labor-intensive nature of sequential extraction experiments, the heterogeneous nature of soil properties, and the interplay of numerous confounding factors (detailed in Supplementary Note 1).” (Page 4, line 71-73)

4. L291, the word ‘Should’ can be replaced by ‘might’

Response:

Suggestion taken. We have revised the sentence accordingly.

5. In L401 and 402, how can you say, “indicating its high generalization capability”? Is it considering the R2 values? If so, can you please put a reference to support this statement? Or if anything else, please elaborate on that.

Response:

We thank the reviewer for this remark. The statement is indeed based on the R^2 value. We have modified the sentence to clarify it, with several supporting references from leading journals provided.

Revised Statement:

“Although the optimal XGBoost model was not trained on any measurements from the independent test dataset, it nevertheless achieved an R^2 value of 0.63 (Supplementary Fig. 14); this indicates its ability to predict metal(loid) fractionation of unknown soil samples with accuracy[33,58]. ” (Page 23, line 437-440)

Response to Reviewer #2

(Manuscript Number: NCOMMS-24-75613)

Overview:

The manuscript reports the development of a machine learning (ML) model to predict the potential for metal(loid) mobility globally, with outcomes presented in terms of both fraction of metal(loid)s likely to be mobile and, where metal(loid) concentrations are available, the absolute mobile concentrations expected. The scope and scale of the work is considerable, for which the authors are to be commended. The work is also presented and communicated well. I am not aware of anything similar having been attempted previously. In this review I have focused primarily on the purpose, applicability and value of the model developed – I have only very limited experience of ML and allied disciplines and so am not able to review much of the technical methodology with any confidence. I have the following questions and comments:

Response:

We thank the reviewer for the positive comments on our manuscript, and for highlighting that our study presents the first attempt at global assessment of soil metal(loid)s fractionation. We also thank the reviewer for the valuable suggestions that have helped us to improve our manuscript, as explained below. Below are point-to-point responses in which we explicitly refer to the relevant lines in the revised manuscript/supplementary information. The comments appear in black, our responses appear in blue, and the texts quoted from the revised manuscript/supplementary information appear in red.

Specific Comments:

1. Perhaps my most significant question regards the potential application of the model. Why is it necessary to obtain comprehensive and fine-resolution data at continental and global scales (Lines 62-64), and what purpose would this have? I can see that there may be value at a very high level in discussions about land use and planning, and there is speculation on this in lines 434-438 (albeit somewhat vague), but at the site scale this would likely only act as an indicator at best. I would not expect it to provide sufficient information or confidence to support local level or site-specific management or remediation as it is highly likely there will be substantial variation in data within the 5x5km resolution (e.g. variability in local geology, soil amendments, impact of site development) that will affect soil properties including carbon levels and pH. However, with site-specific data (including metal(loid) concentrations) this could have value in acting as an alternative to carrying out sequential extractions. The purpose and value of the work should be clarified considerably.

Response:

We thank the reviewer for providing this valuable feedback, and we agree that the trained ML models have different implications at the global and site scales. We have presented a detailed explanation of these implications in the revised manuscript, which makes its significance clear at different scales. In brief, soil metal(loid)s fractionation studies at different scales are all necessary and have different real-world implications. The large-scale studies could shed light on agricultural management, industry planning, and policy-making, while site-scale studies could help with precise pollution evaluation and remediation.

Our study addresses a critical gap in the literature by presenting the application of machine learning to predict global soil metal(loid) mobility, as positively-commented by the all reviewers. The combination of geoscience data and statistical learning techniques to generate large-scale predictive maps has been a widely-adopted framework, such as in arsenic contamination in Chinese soils (*Nature Sustainability, 2024, 7, 766-775*) and in prediction of arsenic in global groundwater (*Science, 2020, 368, 845-850*). Compared to site-scale studies, such global-scale assessments have unique and significant implications. For example, a global map of soil metal(loid)s fractionation will guide policymakers on agricultural and industry planning; prevent further pollution by controlling pollution sources; and reduce risks to public health and the environment [1]. The presented metal(loid) fractionation maps could raise awareness for regions with limited soil fractionation data, particularly for low- and middle-income countries with extensive crop production (i.e., southern Brazil and northern Argentina), and should be used as a guide to further soil metal(loid) fractionation testing [2, 3].

Our trained ML model can also help with site-specific studies related to soil metal(loid)s fractionation. We agree with the reviewer that though our predictive maps have a widely-acceptable resolution for global maps (5-km resolution in our manuscript, compared to 0.5° in *Science, 2024, 384, 233-239* or 10-km in *Nature Geoscience, 2021. 14, 206-210*), there must be substantial variation in soil data within the 5x5km grid. Thus, the generated global maps cannot be used directly to guide site-specific assessment and remediation. Instead, as mentioned by the reviewer, our trained ML models can be used to predict soil metal(loid) fractionation when detailed soil information is obtained at specific sites, substuting, or at least partially, sequential extraction experiments. Therefore, the trained ML models are expected to significantly reduce the resources and time required for site-scale assessment and remediation.

Revised Statement:

"The metal(loid) mobility model can be used as a guide for prioritizing metal(loid) mobility testing and for raising awareness, for example, in central and northern Asia, northern parts of North America and Africa, southern Oceania, and South America. Prioritizing further metal(loid) mobility testing holds more importance for low- and middle-income countries with extensive crop production (i.e., southern Brazil and northern Argentina), where the current metal(loid) mobility condition is largely unknown and limited testing resources must be employed effectively in the future. We note that the presented soil metal(loid) mobility maps cannot be used to guide site-scale assessment and remediation under the current resolution (5 km $\times$ 5 km), particularly because site-scale (< 100 m) heterogeneous soil metal(loid)s and properties cannot be obtained from exiting global datasets[32]. Nevertheless, our soil metal(loid) mobility model can continue to support site-scale soil metal(loid) assessment and remediation by substituting, at least partially, costly and time-consuming sequential extraction experiments. Thus, embracing our model in site-scale studies can provide practitioners with a unique ability to assess soil metal(loid) mobility rapidly and reliably." (Page 15, line 266-280)

2. It is somewhat disingenuous to say that the study has identified significant risks, when this study doesn't appear to consider the actual risk to human health or other factors. The assessment of the potential mobility of contaminants is part of the risk analysis, but not the whole. The mobility may be significant if contamination were to occur, at levels sufficient to cause harm and if receptors (assuming a typical source-pathway-receptor risk analysis approach) are present. The analysis comes up with very high proportions (e.g. 82%) of land at risk from metal(loid) mobility, but this doesn't mean the metal(loids) are actually there and that a significant risk would arise even if they were. I think some statements are potentially misleading – for example, the caption for Figure 2 states that this shows "Risk hotspots of prevalent metal(loid)s" – this is not the case, as the mobility of contaminants is only part of the risk analysis. I'm of the opinion that there is a high level of value in the work done, though the exact nature of this value needs to be clarified accurately throughout.

Response:

We thank the reviewer for rasing the issue of our use of the term 'risk' throughout the manuscript. Indeed, we agree with the reviewer that the risk assessment in the manuscript cannot represent the actual soil metal(loid) risk to the environment. We assume the same total metal(loid) content throughout the world, and investigate soil metal(loid) mobility under these senarios. Though such scenario analysis is widely-acceptable in leading journals (i.e., *Nature, 2022, 608, 719-723*, *Nature Food, 2020, 1, 166-172*) and can provide valuable insights, it cannot represent the real-world metal(loid) levels. In addition, our study primarily focuses on soil metal(loid) mobility, and doesn't consider the complete source-pathway-receptor.

To avoid any ambiguity, we have defined and constantly used 'mobilization risk' throughout the revised manuscript. The mobility analysis in the current study indicates the spillover risk of soil metal(loid) contamination to water and food security, and its associated risk is termed as mobilization risk in the current study. Accordingly, we have made modifications in all necessary parts in the revised manuscript, such as title, figure title, and main text. Moreover, we acknowledge that assessing the actual risk involving the source-pathway-receptor framework is much more complex and clearly out of the current scope. We have also clearly explained this limitation in our discussion part.

We also thank the reviewer for the positive comments on the high level of value in our manuscript.

Revised Statement:

"Global and regional patterns of soil metal(loid) mobility and associated risks" (Page 1, line 1-2)

"The baseline mobility measures the most mobile metal(loid) content once baseline contamination occurs; a high baseline mobility represents a high likelihood of a soil metal(loid) being mobilized into soil pore water. Thus, the mobility analysis in the present study indicates the spillover risk of soil metal(loid) contamination to water and food security, and its associated risk is termed as the mobilization risk in subsequent analyses." (Page 5, line 103-108)

"**Fig. 2 Global mobilization risk hotspots in 2024, revealed by the soil metal(loid) mobility model. a,** Mobilization risk hotpots of prevalent metal(loid)s (Pb, As, Cd, Cr, Cu, and Hg) across the globe. This plot illustrates the number of prevalent metal(loid)s identified with medium-to-high mobilization risk in baseline scenarios for a specific site; **b–i,** Zoomed-in sections of the typical mobilization risk hotspots." (Page 10, line 174-178)

"We used the risk assessment code to evaluate the mobilization risk of metal(loid)s in the soil. Specifically, low and medium-to-high mobilization risks were associated with a P_{F1} of 1–10% and >10%, respectively [67]." (Page 25, line 483-485)

"We also acknowledge that our metal(loid) mobility model primarily focuses on the mobility of soil metal(loid)s; therefore, the mobilization risk analysis could shed light on the likelihood of a metal(loid) in soil being mobilized under certain contamination scenarios. In other words, the global mobility maps of soil metal(loid)s cannot represent their actual risk, since the global distribution of the soil metal(loid) content is unavailable, and the source–pathway–receptor interactions are not considered." (Page 18, line 337-343)

3. The data used for training and validation of the model isn't uniform globally or within countries. There is quite a good level of coverage, but there are gaps (to be expected). However, are the training data sufficiently representative of global conditions that the resulting model is accurate everywhere? The model is validated on a subset of the data used, but given the validation datasets in particular appear to be limited to particular geographical regions (Supplementary Table 8 suggests a large proportion is from China, whilst all the lab tests are from China) then is this valid for other parts of the world?

Response:

We agree with the reviewer that there are gaps in the spatial coverage of the current global dataset. However, we want to make it clear that the model developed in this study does not use spatial location (or geographic coordinates) as a predictor. Consequently, spatial generalization of the model is driven by similarity in the environmental and element-related variables (e.g., soil properties), rather than by spatial proximity (as geostatistics does). We also added an additional analysis of the degree of interpolation/extrapolation in the revised manuscript, following the approach of van den Hoogen, Geisen [4]. It is confirmed that our model required very limited extrapolation to environmental conditions not included in the training plots, manifesting the reliability of its global application.

In order to showcase the applicability of trained ML model on unknown samples, extensive validations have been performed in our revised manuscript, including:

- ✓ We used the conventional validation method in ML that splits the whole dataset into training and test sets. Soil samples in the test set are unknown to the trained ML models. The optimal XGBoost model could establish a strong correspondence between the observed and predicted fractionation values on the test set (R^2 = 0.86, MAE = 6.21), indicating a good predictive accuracy on unknown soil samples (for example, an R^2 = 0.71 was achieved on the validation set in *Nature Sustainability, 2024, 7, 766-775*).

- ✓ To further assess the out-of-sample error, we compiled an independent test dataset from the most recent studies and extensive laboratory experiments. Although the optimal XGBoost was not trained on any measurements from the external dataset, the model nevertheless achieved an R^2 of 0.63. Considering the reviewer's comments, we have screened more than 1,000 published papers to include more data from other parts of the world. Unfortunely, we can only add another 468 data points from other parts of the world (except China) due to data availability. The optimal model still performed well on the expanded independent test dataset.

- ✓ We have also added a hold-out test to indicate the generalization capability of our ML gramework to other regions. The idea of this hold-out test also originates from the reviewer's comment on EU-focused analysis. In this hold-out test, we used all EU data as the test data points and the data from other parts of the world as the training data points. In doing so, the XGBoost model was not trained on any EU data during its construction. We found that the trained XGBoost model achieved an R^2 of 0.50 in the hold-out test, indicating a good generalization capability to other regions around the world (for example, an R^2 of 0.445 was achieved on the idependent test set in *Nature, 2020, 585, 545-550*).

- ✓ To further ensure the reliability of the model in real-world applications, we employed a series of uncertainty analyses to demonstrate the trustworthy of trained ML models in the revised manuscript. Our uncertainty analyses implied the high reliability in the model prediction results and good generalization of models.

Revised Statement:

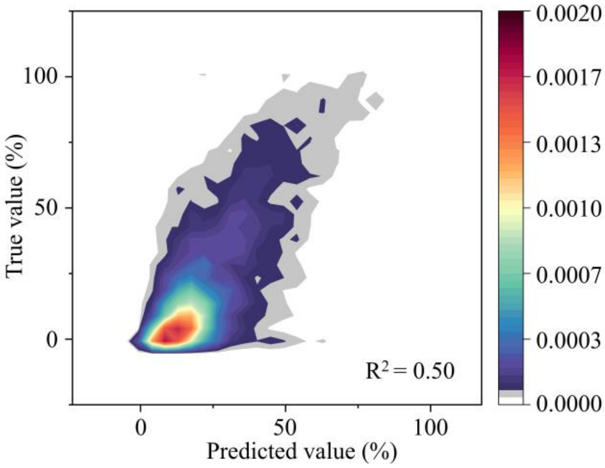
"**Modelling reliability, prediction uncertainty, and descriptor importance**

In addition to training-test verification, we employed a series of reliability analyses to assess the out-of-sample error of our model. We compiled an independent test dataset from the most recent studies (1,620 data points, Supplementary Table 8) and extensive laboratory experiments (3,748 data points). The compiled independent test dataset primarily contained soil samples from China and EU (details in Supplementary Note 8). Although the optimal XGBoost model was not trained on any measurements from the independent test dataset, it nevertheless achieved an R^2 value of 0.63 (Supplementary Fig. 14); this indicates its ability to predict metal(loid) fractionation of unknown soil samples with accuracy[33,58]. We also performed a hold-out test to showcase that the trained XGBoost model had a good generalization capability for other regions that were not covered during model training (Supplementary Note 9).

In the present study, we utilized resampling methods based on the jackknife strategy to estimate robust prediction intervals with the optimal XGBoost model[59]. Ideally, the uncertainty quantification of the optimal XGBoost model should yield a c% confidence interval that contains the true value for approximately c% of the time[60]. Our uncertainty analysis revealed that 95.49% of the testing data fell into the 95% confidence intervals of prediction, indicating low uncertainty in predicting soil metal(loid) fractionation during testing." (Page 23, line 443-449)

"S9: Hold-out test of the XGBoost model.

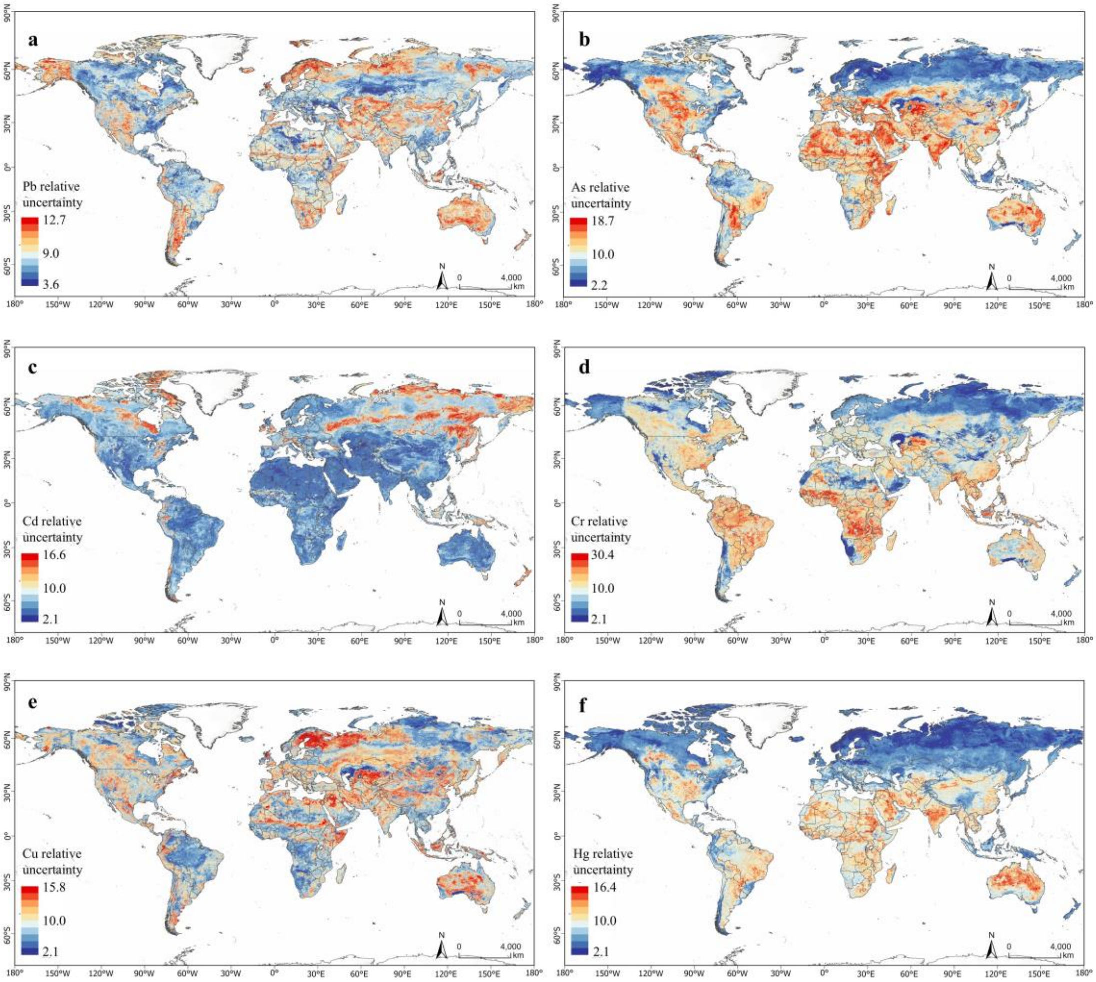
Since the XGBoost model was used to spatially map the metal(loid) mobility in the soil of EU member states in 2009, we performed a hold-out test of the trained XGBoost model on EU soil samples. To do so, we used all data points from EU (12,260 data points) as the test data and data points from other parts of the world (18,572 data points) as the training data. The XGBoost algorithm with the optimal hyperparameters was re-trained using the new training data and its performance was validated on the EU test data. We found that the trained XGBoost model achieved an R^2 of 0.50 on the EU test data (Supplementary Fig. 25), though it was not trained on any EU data during its construction. The above hold-out test confirmed the good generalization capability of the trained ML models to other regions around the world.



Supplementary Fig. 25 Scatter plot of true and predicted output on the hold-out EU test data." (Supplementary Note 9)

"The relative uncertainty maps of global prediction were constructed by displaying standard deviation divided by the mean of prediction based on 20 runs of our final XGBoost model (Supplementary Fig.

15). The relative uncertainty of global mobile fractions for prevalent metal(loid)s in baseline scenarios (Fig. 1a－f) was averaged at 9.70%, indicating a small relative uncertainty in the global application of the model[66]." (Page 24, line 473-478)



Supplementary Fig. 15 Distribution of the relative uncertainty corresponding to mobile fraction for prevalent metal(loid)s in baseline scenarios.

"Before the global application of the optimal XGBoost model, the degree of interpolation/extrapolation was performed[62]. It is found that the optimal XGBoost model required very limited extrapolation to soil properties not included in the training plots (0.13% of soil OC and 0% of other soil properties, Supplementary Table 9), manifesting the reliability of its global application." (Page 24, line 461-465)

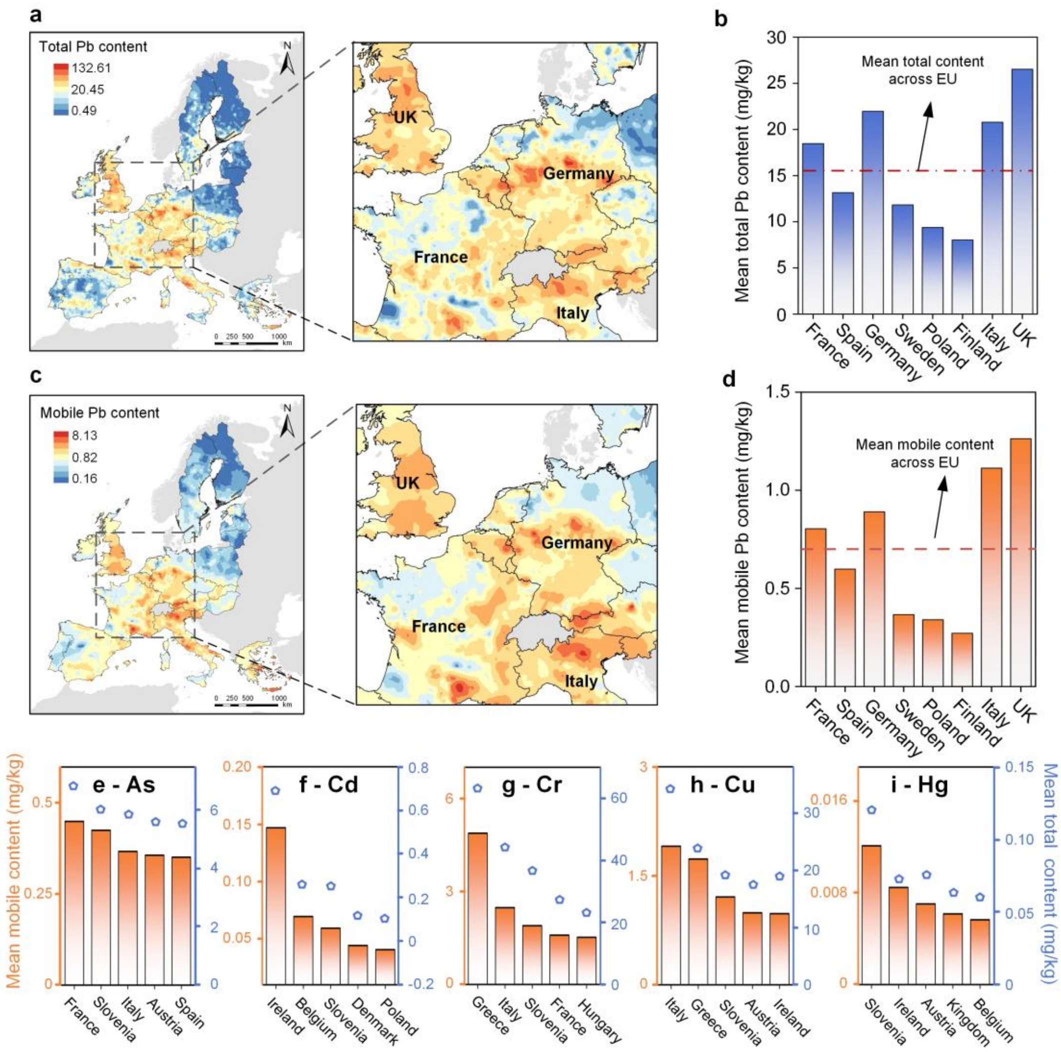
4. Spatial disparities in EU soil (lines 207-243): In Figure 4, comparing the total Pb content in France and the UK there looks (visually at least) to be significant difference in content from the map (Fig 4a) whilst Fig 4b suggests there is only a small difference in mean content. The majority of the UK, for example, appears to be at the highest level of Pb (coloured orange, appears >100 mg/kg?) whilst France is much more variable. Depending on where the UK RGVs lie (see next comment), this may also imply that the vast majority of the UK is above typical RGVs in EU countries – is this really true? Does this result from the colour scale of lead content used being apparently non-

uniform, with 40-50% of the Pb range considered being coloured orange? Unless there is good reason for this (e.g. if it relates to the RGVs in these regions somehow), perhaps the colour spectrum used should vary more linearly to more clearly distinguish the levels of contamination.

Response:

We thank the reviewer for providing this valuable feedback. We comfirm that the statistics are correct and the misunderstanding is indeed caused by the colour scale in the distribution maps. We have re-plotted Figure 4 and Supplementary Figures. After comparison, we prefer to use geometric interval classification in the color spectrum, which is widely used in leading journals (such as *Nature Geoscience, 2021. 14, 206-210*). The geometric interval algorithm creates geometric intervals by minimizing the sum of squares of the number of elements in each class. This method typically provides results that are cartographically comprehensive, highlighting neither the middle nor extreme ranges of values.

Revised Statement:



"

Fig. 4 Continental prediction of soil metal(loid) mobility at the European scale. a, Distribution of total Pb content, along with zoomed-in sections in central Europe; **b**, Mean total Pb content of European countries with soil samples >500; **c**, Modeled distribution of mobile Pb content, along with zoomed-in sections in central Europe; **d**, Modeled mean mobile Pb content of European countries with soil samples >500; The total and mobile contents in maps are classified using the geometric interval method in the color spectrum. **e-i**, Mean total and mobile contents of As, Cd, Cr, Cu, and Hg for the top five EU countries. The orange bars on the left-hand side y-axis represent the mobile metal(loid) content, and the blue markers on the right-hand side y-axis represent the total metal(loid) content. The countries are ordered based on their mobile metal(loid) content. " (Page 15, line 248-257)

5. Line 216 - it is stated that the majority of RGVs for Pb are 50 or 100 mg/kg, yet in Supplementary Table 2 the baseline RGV is 150 mg/kg for the EU. Can this be explained and/or corrected?

Response:

We thank the reviewer for this comment. To avoid ambiguity, modifications have been made in the main text and supplementary table. In the main text, we changed contamination expression to high levels of metal(loid) content, since different RGVs are used by different EU members states (the RGVs widely adopted are 50 mg/kg, 100 mg/kg, and 200 mg/kg). The baseline RGV for Pb was selected since 150 mg/kg was close to the geometric mean.

Revised Statement:

"For example, the total Pb content ranged between 1.99 and 151.12 mg/kg, with an average of 15.07 mg/kg. High levels of Pb contents were observed around central United Kingdom (UK), central Germany, and central Italy." (Page 13, line 228-231)

"**Supplementary Table. 2 The regulatory guidance values (RGVs) for baseline and elevated contamination scenarios of prevalent metal(loid)s.**

Metal(loid)s	Scenarios	
	RGVs (mg/kg) Baseline	RGVs (mg/kg) Elevated contamination
Cd	1	70
As	0.39	20
Cu	100	3100
Hg	1	10
Pb	150	400
Cr	3200	120000

Note: The RGVs come from multinational organizations, national regulations or close to geometric mean values of most RGVs."

6. Is there any external validation of the total and/or mobile metal(loid)s contents found in this EU-focused analysis? For example, how does independent literature stemming from the EU compare to the model outputs?

Response:

We thank the reviewer for this remark. The total metal(loid) contents in EU originate from the LUCAS project, in which around 20,000 soil samples were comprehensively analysed using standard analytical procedures in a single laborotory. The LUCAS database has been one of the largest spatial databases of the soil cover across the EU. Extensive soil analysis and modelling studies have been carried out using the LUCAS database (such as *Nature Geoscience*, *2021*, *14*, *295-300*). Thus, the total metal(loid) contents used in the current manuscript are widely recognized to be accurate and reliable.

The mobile metal(loid) contents in EU soils were predicted using the optimal XGBoost model with reliable soil information from LUCAS database as inputs. The reliability and generalization capability of the trained ML models have been well explained within the reply to your third comment (please refer to the above response). We want to mention the following two points again, which are related to the external validation for EU-focused analysis:

- ✓ The independent test dataset contained independent literature stemming from the EU. As mentioned above, our optimal XGBoost model performed well on the independent test dataset (R^2 = 0.63), indicating its generalization capability in EU soils.

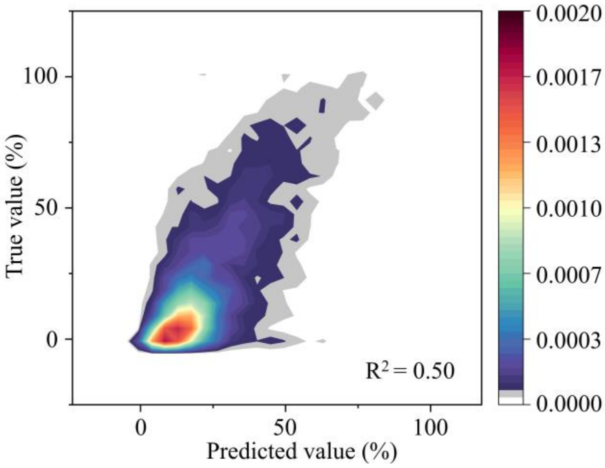
✓ We have also added a hold-out test to indicate the generalization capability of our ML gramework to EU. In this hold-out test, we used all EU data as the test data points and the data from other parts of the world as the training data points. In doing so, the XGBoost model was not trained on any EU data during its construction. We found that the trained XGBoost model achieved an R^2 of 0.50 in the hold-out test, indicating a good generalization capability to EU soils (for example, an R^2 of 0.445 was achieved on the idependent test set in *Nature, 2020, 585, 545-550*).

Revised Statement:

"In addition to training-test verification, we employed a series of reliability analyses to assess the out-of-sample error of our model. We compiled an independent test dataset from the most recent studies (1,620 data points, Supplementary Table 8) and extensive laboratory experiments (3,748 data points). The compiled independent test dataset primarily contained soil samples from China and EU (details in Supplementary Note 8). Although the optimal XGBoost model was not trained on any measurements from the independent test dataset, it nevertheless achieved an R^2 value of 0.63 (Supplementary Fig. 14); this indicates its ability to predict metal(loid) fractionation of unknown soil samples with accuracy[33,58]. We also performed a hold-out test to showcase that the trained XGBoost model had a good generalization capability for other regions that were not covered during model training (Supplementary Note 9)." (Page 23, line 432-442)

"S9: Hold-out test of the XGBoost model.

Since the XGBoost model was used to spatially map the metal(loid) mobility in the soil of EU member states in 2009, we performed a hold-out test of the trained XGBoost model on EU soil samples. To do so, we used all data points from EU (12,260 data points) as the test data and data points from other parts of the world (18,572 data points) as the training data. The XGBoost algorithm with the optimal hyperparameters was re-trained using the new training data and its performance was validated on the EU test data. We found that the trained XGBoost model achieved an R^2 of 0.50 on the EU test data (Supplementary Fig. 25), though it was not trained on any EU data during its construction. The above hold-out test confirmed the good generalization capability of the trained ML models to other regions around the world.



Supplementary Fig. 25 Scatter plot of true and predicted output on the hold-out EU test data.” (Supplementary Note 9)

Response to Reviewer #3

(Manuscript Number: NCOMMS-24-75613)

Overview:

This manuscript presents a novel application of machine learning (ML) for modeling the global mobility of soil metal(loid)s. From the perspective of a soil scientist, this study marks the first attempt to investigate soil metal(loid)s mobility on a global scale, leading to several intriguing findings. The work is of significant interest to a broad audience and is suitable for publication in Nature Communications. However, I propose the following suggestions for improvement before its acceptance.

Response:

We thank the reviewer for the positive appreciation of our manuscript. Revision based on your comments has been made to further improve the quality of this paper. Below are point-to-point responses in which we explicitly refer to the relevant lines in the revised manuscript/supplementary information. The comments appear in black, our responses appear in blue, and the texts quoted from the revised manuscript/supplementary information appear in red.

Specific Comments:

1. The authors should elaborate on the rationale for selecting the BCR sequential extraction method. Does this method adequately represent metal(loid)s mobility? Please provide further discussion.

Response:

We thank the reviewer for this comment. We have added more explanations and comments on the selection of BCR and whether it can represent metal(loid) mobility in the revised manuscript. In brief, sequential extraction has been widely used to fractionate soil metal(loid)s into operationally defined forms according to their physiochemical mobility and potential availability [5]. It has been widely used in environmental analytical chemistry to represent metal(loid) mobility [6], and can work as a common scheme so that results from different researchers can be compared and an international database can be established [7, 8]. Among the widely-adopted sequential extraction techniques, BCR was selected due to its simple implementation, repeatable fractionation results, and data availability [9].

Revised Statement:

"Another notable limitation arises from the use of metal(loid) fractionation to represent mobility. Sequential extraction has been widely used to fractionate soil metal(loid)s into operationally defined forms according to their physiochemical mobility and potential availability [34]. This method has been proven to be useful in the field of environmental analytical chemistry [35] and can work as a common scheme so that results from different researchers can be compared and an international database can be established [36,37]." (Page 18, line 326-332)

"Considering the simple implementation, repeatable fractionation results, and data availability[40], we focused on studies involving metal(loid) fractionation from the Community Bureau of Reference (BCR) four-step sequential extraction (N≈1,000 studies)." (Page 19, line 353-356)

2. What exactly does metal(loid)s mobility entail? How is this mobility connected to the actual risk of metal(loid)s pollution? The authors should clarify this linkage.

Response:

We thank the reviewer for this comment and for the opportunity to clarify the linkage between mobility and actual risk. The mobility analysis in the current study indicates the spillover risk of soil metal(loid) contamination to water and food security, and its associated risk is termed as mobilization risk in the present study. We admit that the mobility and mobilization risk analysis in the current study cannot represent the actual soil metal(loid) risk to the environment, but it is a key step in evaluating the actual risk. We assumed the same total metal(loid) content throughout the world, and investigate soil metal(loid) mobility under these senarios. Though such scenario analysis is widely-acceptable in leading journals (i.e., *Nature, 2022, 608, 719-723, Nature Food, 2020, 1, 166-172*) and can provide valuable insights, it cannot represent the real-world metal(loid) levels. In addition, our study primarily focuses on soil metal(loid) mobility, and doesn't consider the complete source-pathway-receptor.

Accordingly, we have made modifications in all necessary parts in the revised manuscript, such as title, figure titles, and main text. Moreover, failure to address the actual risk is a limitation of the current study, though the actual risk is much more complex and clearly out of the current scope. We have also clearly explained this limitation in the discussion part.

Revised Statement:

"The baseline mobility measures the most mobile metal(loid) content once baseline contamination occurs; a high baseline mobility represents a high likelihood of a soil metal(loid) being mobilized into soil pore water. Thus, the mobility analysis in the present study indicates the spillover risk of soil metal(loid) contamination to water and food security, and its associated risk is termed as the mobilization risk in subsequent analyses." (Page 5, line 103-108)

"We also acknowledge that our metal(loid) mobility model primarily focuses on the mobility of soil metal(loid)s; therefore, the mobilization risk analysis could shed light on the likelihood of a metal(loid) in soil being mobilized under certain contamination scenarios. In other words, the global mobility maps of soil metal(loid)s cannot represent their actual risk, since the global distribution of the soil metal(loid) content is unavailable, and the source–pathway–receptor interactions are not considered." (Page 18, line 337-343)

3. The global application of the trained ML model is predicated on the assumption of consistent total metal(loid)s content. While this is a widely accepted approach in the absence of global distribution

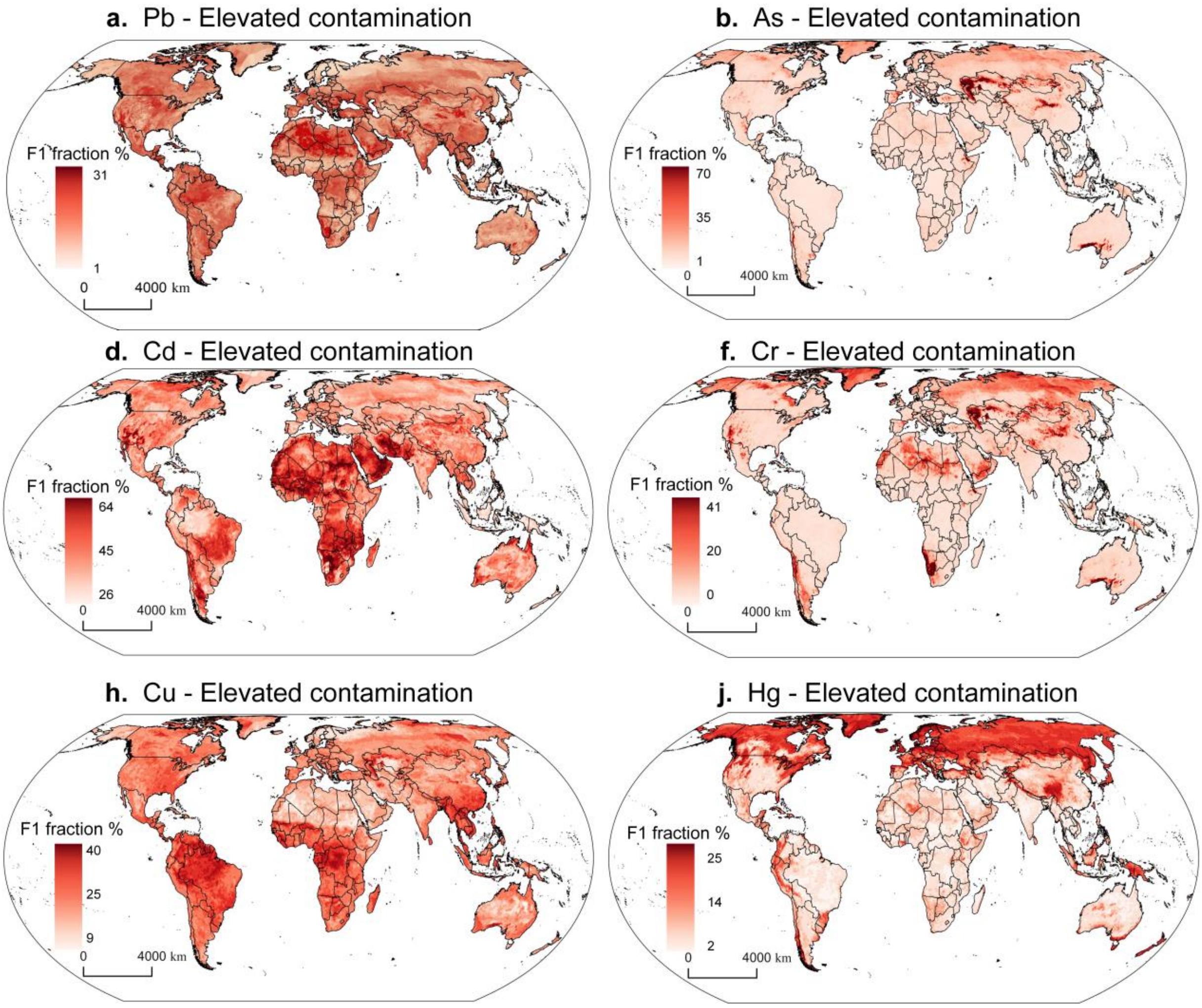
data, it raises an important question: How might changes in total metal(loid)s content impact the global distribution, particularly the identification of hotspots? This warrants further investigation.

Response:

We thank the reviewer for this valuable suggestion. We agree with the reviewer that scenario analysis is widely used in leading journals when some key information for model application is missing (i.e., *Nature, 2022, 608, 719-723*, *Nature Food, 2020, 1, 166-172*). Following the reviewer's comment, our metal(loid) mobility model revealed that the change in total metal(loid) content will influence the mobility distribution. Evident variations in the increase of mobile metal(loid)s fractions and contents were observed from baseline to elevated contamination scenarios (Supplementary Figure 3), indicating that soil properties had a crucial influence on the mobility distribution and dynamics across the globe. We have provided more detailed results and discussion about the influence of total metal(loid) content in the revised manuscript.

Revised Statement:

"Notably, the mobilization risk evidently increased with the contamination level. Given the elevated contamination levels (Methods), the global average land fraction with P_{F1} > 10% for the most prevalent metal(loid)s would expand to 52.73% (Supplementary Table 1), as predicted by the ML model. Evident variations in the increase of the mobile fractions were observed from baseline to elevated contamination scenarios (Supplementary Fig. 3), indicating that soil properties had a crucial influence on the mobility distribution and dynamics across the globe. For example, the mobile Pb fraction increased more evidently in South America and Africa than in other regions under the same increase in total Pb content. However, the mobilization risk hotspots remained similar under both the baseline and elevated contamination scenarios." (Page 6, line 123-132)



Supplementary Fig. 3 Global prediction of mobile fractions of prevalent metal(loid)s in soil at elevated contamination scenarios.

4. The uncertainty analysis should be conducted for the ML modeling.

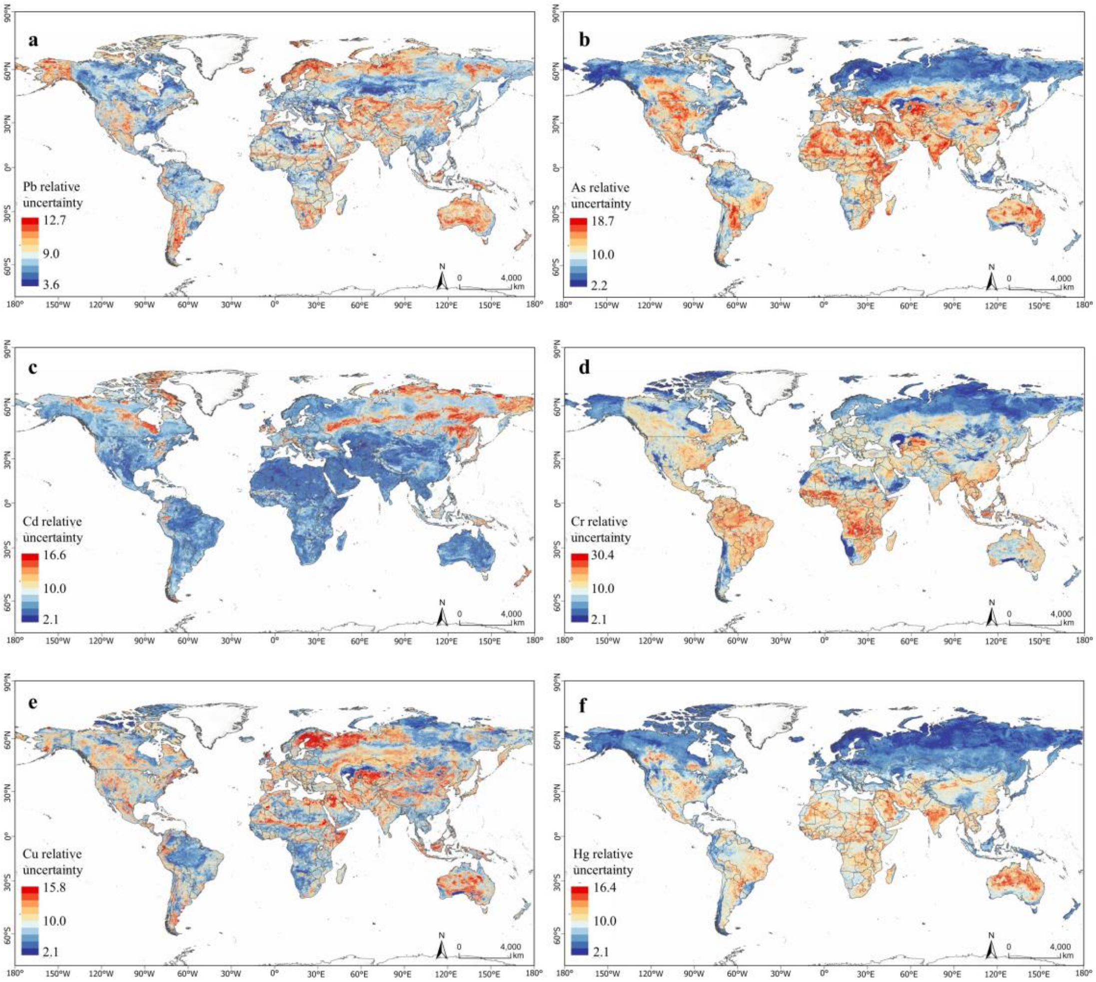
Response:

We thank the reviewer for this valuable suggestion. Considering this, we employed a series of uncertainty analyses to demonstrate the trustworthy of trained ML models in the revised manuscript. Uncertainty analyses of the optimal model were conducted during testing (using prediction confidence intervals) and model application (using relative uncertainty maps). Our uncertainty analyses implied the high reliability in the model prediction results and good generalization of models.

Revised Statement:

"In the present study, we utilized resampling methods based on the jackknife strategy to estimate robust prediction intervals with the optimal XGBoost model[59]. Ideally, the uncertainty quantification of the optimal XGBoost model should yield a c% confidence interval that contains the true value for approximately c% of the time[60]. Our uncertainty analysis revealed that 95.49% of the testing data fell into the 95% confidence intervals of prediction, indicating low uncertainty in predicting soil metal(loid) fractionation during testing..” (Page 23, line 443-449)

“The relative uncertainty maps of global prediction were constructed by displaying standard deviation divided by the mean of prediction based on 20 runs of our final XGBoost model (Supplementary Fig. 15). The relative uncertainty of global mobile fractions for prevalent metal(loid)s in baseline scenarios (Fig. 1a－f) was averaged at 9.70%, indicating a small relative uncertainty in the global application of the model[66].” (Page 24, line 473-478)



Supplementary Fig. 15 Distribution of the relative uncertainty corresponding to mobile fraction for prevalent metal(loid)s in baseline scenarios.

5. The reliability of the ML model for global applications should be addressed. Can the model consistently provide stable predictions across diverse regions?

<u>Response:</u>

We thank the reviewer for this comment and for the opportunity to perform more reliability analyses. The reliability of the trained ML was conducted in multiple ways in the revised manuscript, including:

- ✓ We used the conventional validation method in ML that splits the whole dataset into training and test sets. Soil samples in the test set are unknown to the trained ML models. The optimal

XGBoost model could establish a strong correspondence between the observed and predicted fractionation values on the test set (R^2 = 0.86, MAE = 6.21), indicating a good predictive accuracy on unknown soil samples (for example, an R^2 = 0.71 was achieved on the validation set in *Nature Sustainability, 2024, 7, 766-775*).

- ✓ To further assess the out-of-sample error, we complied an independent test dataset from the most recent studies and extensive laboratory experiments. Although the optimal XGBoost was not trained on any measurements from the external dataset, the model nevertheless achieved an R^2 of 0.63.

- ✓ We have also added a hold-out test to indicate the generalization capability of our ML gramework to other regions. The idea of this hold-out test also comes from the reviewer's comment on EU-focused analysis. In this hold-out test, we used all EU data as the test data points and the data from other parts of the world as the training data points. In doing so, the XGBoost model was not trained on any EU data during its construction. We found that the trained XGBoost model achieved an R2 of 0.50 in the hold-out test, indicating a good generalization capability to other regions around the world (for example, an R^2 of 0.445 was achieved on the idependent test set in *Nature, 2020, 585, 545-550*).

- ✓ We also added an additional analysis of the degree of interpolation/extrapolation in the revised manuscript, following the approach of van den Hoogen, Geisen [4]. It is confirmed that our model required very limited extrapolation to environmental conditions not included in the training plots, manifesting the reliability of its global application.

Together with the uncertainty analyses explained in the fourth comment, the trained ML models can consistently provide accurate and stable predictions across diverse regions.

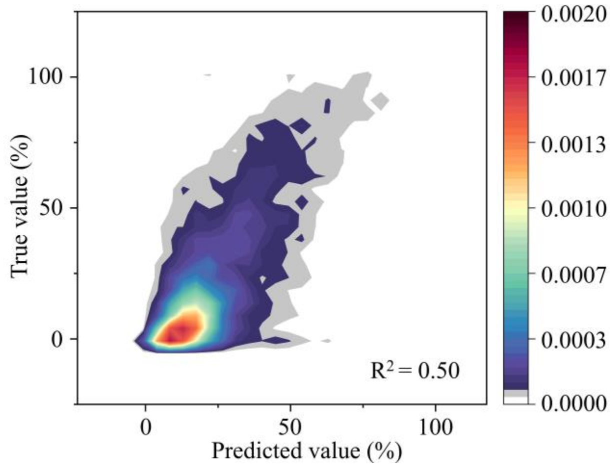
Revised Statement:

"**Modelling reliability, prediction uncertainty, and descriptor importance**

In addition to training-test verification, we employed a series of reliability analyses to assess the out-of-sample error of our model. We compiled an independent test dataset from the most recent studies (1,620 data points, Supplementary Table 8) and extensive laboratory experiments (3,748 data points). The compiled independent test dataset primarily contained soil samples from China and EU (details in Supplementary Note 8). Although the optimal XGBoost model was not trained on any measurements from the independent test dataset, it nevertheless achieved an R^2 value of 0.63 (Supplementary Fig. 14); this indicates its ability to predict metal(loid) fractionation of unknown soil samples with accuracy[33,58]. We also performed a hold-out test to showcase that the trained XGBoost model had a good generalization capability for other regions that were not covered during model training (Supplementary Note 9)." (Page 23, line 432-442)

"S9: Hold-out test of the XGBoost model.

Since the XGBoost model was used to spatially map the metal(loid) mobility in the soil of EU member states in 2009, we performed a hold-out test of the trained XGBoost model on EU soil samples. To do so, we used all data points from EU (12,260 data points) as the test data and data points from other parts of the world (18,572 data points) as the training data. The XGBoost algorithm with the optimal hyperparameters was re-trained using the new training data and its performance was validated on the EU test data. We found that the trained XGBoost model achieved an R^2 of 0.50 on the EU test data (Supplementary Fig. 25), though it was not trained on any EU data during its construction. The above hold-out test confirmed the good generalization capability of the trained ML models to other regions around the world.



Supplementary Fig. 25 Scatter plot of true and predicted output on the hold-out EU test data." (Supplementary Note 9)

"Before the global application of the optimal XGBoost model, the degree of interpolation/extrapolation was performed[62]. It is found that the optimal XGBoost model required very limited extrapolation to soil properties not included in the training plots (0.13% of soil OC and 0% of other soil properties, Supplementary Table 9), manifesting the reliability of its global application." (Page 24, line 461-465)

6. Why was the EU chosen for this study? Large-scale metal(loid)s data are also available for other countries, such as China. The authors should justify this choice.

Response:

We thank the reviewer for these comments. We have provided more detailed reasons for the selection of EU in the revised manuscript.

Revised Statement:

"We applied the optimal XGBoost model to spatially map the metal(loid) mobility in the soil of EU member states in 2009. We selected EU member states as a case study for several reasons. First, standard and extensive soil sampling has been conducted in the Land Use and Coverage Area frame

Survey (LUCAS) project, which encompassed a large area with a wide distribution of soil types. Second, comprehensive analyses, including the analyses of soil properties and total contents of prevalent metal(loid)s, have been performed in a single laboratory to ensure consistency within the database. Third, high-quality datasets containing data regarding the required descriptors for predicting metal(loid) mobility have been made publicly available for EU member states via resources offered by international agencies. Finally, in addition to the total metal(loid) content that has been analyzed[68], an analysis of their mobility would provide a more accurate assessment of their mobilization risk." (Page 26, line 511-522)

7. In Lines 419–421, the authors mention selecting 22 metal(loid)s with their RGVs within the data range. Could the authors explain why this selection criterion was applied in the context of the study?

Response:

We thank the reviewer for this remark. The purpose for selecting 22 metal(loid)s with their RGVs within the data range is to increase the reliability of our prediction results. Though advanced ML algorithms are powerful in learning nonlinear relationships within the database, there might be certain bias during extrapolation [10, 11]. Thus, these 22 metal(loid)s with suitable RGVs were selected to avoid bias in ML extrapolation and ensure the reliability of our prediction.

Revised Statement:

"To avoid bias in ML extrapolation and ensure the reliability of our prediction, 22 metal(loid)s were analyzed because their selected RGVs were within the TMC range of the global dataset[64,65]." (Page 24, line 470-472)

References

1. Hou D, Ok YS. Soil pollution--speed up global mapping. *Nature* **566**, 455-456 (2019).

2. Zhang S*, et al.* Escalating arsenic contamination throughout Chinese soils. *Nature Sustainability*, (2024).

3. Podgorski J, Berg M. Global threat of arsenic in groundwater. *Science* **368**, 845-850 (2020).

4. van den Hoogen J*, et al.* Soil nematode abundance and functional group composition at a global scale. *Nature* **572**, 194-198 (2019).

5. Hlavay J, Prohaska T, Weisz M, Wenzel WW, Stingeder GJ. Determination of trace elements bound to soil and sediment fractions (IUPAC Technical Report). *Pure and Applied chemistry* **76**, 415-442 (2004).

6. A.Tessier, P.G.C.Campbell. Comment on "Pitfalls of sequential extractions" by P. M. V. Nirel and F. M. M. Morel. *Water Research* **25**, 115-117 (1991).

7. Sutherland RA. BCR®-701: A review of 10-years of sequential extraction analyses. *Analytica Chimica Acta* **680**, 10-20 (2010).

8. Tack FMG, Verloo MG. Chemical Speciation and Fractionation in Soil and Sediment Heavy Metal Analysis: A Review. *International Journal of Environmental Analytical Chemistry* **59**, 225-238 (1995).

9. Hernández-Moreno JM, Rodríguez-González JI, Espino-Mesa M. Evaluation of the BCR sequential extraction for trace elements in European reference volcanic soils. *European Journal of Soil Science* **58**, 419-430 (2007).

10. Zhan J*, et al.* Evaluating Interpolation and Extrapolation Performance of Neural Retrieval Models. *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, (2022).

11. Muckley ES, Saal JE, Meredig B, Roper CS, Martin JH. Interpretable models for extrapolation in scientific machine learning. *Digital Discovery* **2**, 1425-1435 (2023).

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

All the raised comments have been addressed. Based on the other reviewers' comments it can be accepted for publication.

Reviewer #2 (Remarks to the Author):

The authors have done well to comprehensively address my questions and concerns with clarity. In specific response to certain rebuttals:

• Point 1: My question on the potential application of the model has been well addressed by the new addition in the discussion. I am of the opinion that this could be slightly enhanced in the abstract and introduction, but this is only a minor recommendation and I am, overall, happy with the response.

• Point 2: The redefinition of risk as mobilisation risk has worked well and seems reasonable. This has been addressed consistently throughout the manuscript and the revised text is much clearer on this point.

• Point 3: The clarification on ML training and validation seems to be reasonable. As this isn't my speciality, then if the other reviewers are happy with this then this will be acceptable to me.

Minor comment:

• Line 275 – "exiting" should be 'existing'.

Reviewer #3 (Remarks to the Author):

I have carefully read the author's responses to the reviewers and the revised manuscript. The authors have madea satisfactory revision on the manuscript and addressed all my comments and suggestions. I am satisfied with the responses and revisions, so I suggest accepting the manuscript for publication in NC.

Response to Reviewer #1

(Manuscript Number: NCOMMS-24-75613A)

Overview:

All the raised comments have been addressed. Based on the other reviewers' comments it can be accepted for publication.

Response:

We thank the reviewer again for your valuable comments and the positive appreciation of our revision.

Response to Reviewer #2

(Manuscript Number: NCOMMS-24-75613A)

Overview:

The authors have done well to comprehensively address my questions and concerns with clarity. In specific response to certain rebuttals:

Response:

We thank the reviewer again for your valuable comments and the positive appreciation of our revision. Further minor revision has been made based on your comments/suggestions. Below are point-to-point responses in which we explicitly refer to the revised content in the revised manuscript/supplementary information. The comments appear in black, our responses appear in blue, and the texts quoted from the revised manuscript/supplementary information appear in red.

Specific Comments:

1. My question on the potential application of the model has been well addressed by the new addition in the discussion. I am of the opinion that this could be slightly enhanced in the abstract and introduction, but this is only a minor recommendation and I am, overall, happy with the response.

Response:

We thank the reviewer for providing this valuable feedback. Accordingly, we have further improved our manuscript by indicating the potential application of the model in the abstract and introduction.

Revised Statement:

"These findings offer crucial insights into global distributions and drivers of soil metal(loid) mobility, providing a robust tool for prioritizing metal(loid) mobility testing, raising awareness, and informing sustainable soil management practices."

"This study highlights the dynamic nature of metal(loid) mobility, providing insights for mitigating global soil pollution and supporting sustainable soil management. The present model can also be used to aid soil metal(loid) assessment, raise awareness, and prioritize further testing of metal(loid) mobility at both local and global scales."

2. The redefinition of risk as mobilisation risk has worked well and seems reasonable. This has been addressed consistently throughout the manuscript and the revised text is much clearer on this point.

Response:

We thank the reviewer again for your previous comment about the use of 'risk' and the positive appreciation of our revision.

3. The clarification on ML training and validation seems to be reasonable. As this isn't my speciality, then if the other reviewers are happy with this then this will be acceptable to me.

Response:

We thank the reviewer again for providing valuable comments/suggestions on the ML training and validation. Once again, the robustness and applicability of our trained ML models have been well validated and verified through multiple ways, including our-of-sample validation, hold-out test, uncertainty analyses, and reliability analyses. Details have been provided in our previous response.

4. Line 275 – "exiting" should be 'existing'.

Response:

Corrected.

Response to Reviewer #3

(Manuscript Number: NCOMMS-24-75613A)

Overview:

I have carefully read the author's responses to the reviewers and the revised manuscript. The authors have madea satisfactory revision on the manuscript and addressed all my comments and suggestions. I am satisfied with the responses and revisions, so I suggest accepting the manuscript for publication in NC.

Response:

We thank the reviewer again for your valuable comments and the positive appreciation of our revision.