Replies to Referee #1's comments on "Insights into the prediction uncertainty of machine-learning-based digital soil mapping through a local attribution approach" (egusphere-2024-323)

We would like to thank Referee #1 for the constructive comments. We agree with most of the suggestions and, therefore, we have modified the manuscript to take on board their comments. We recall the reviews and we reply to each of the comments in turn (outlined in blue). The main corrections made to the manuscript are described in a specific section of each response.

Referee #1:

This is a review for the manuscript Insights into the prediction uncertainty of machine-learning-based digital soil mapping through a local attribution approach by Rohmer et al. The authors use SHAP, a common tool for assessing machine learning predictions at local scale, to investigate the contribution of covariates (or rather groups of covariates) on the uncertainty of a random forest model. It is well known that Shapley values are computationally very expensive, and so the authors propose to reduce the number of covariates to speed up computations. This is done before model training (a rather odd proposal) by using a statistical dependence test (i.e., HSIC), and then after model training by grouping covariates (again with the same dependence test). The main aim of investigating covariates with the model's uncertainty is intriguing within the field of digital soil mapping, but the manuscript has some major flaws. Major concerns are related to the methodology of the entire selection procedure of covariates as well as the with the presented case study. The quality of the writing is also unfortunately poor.

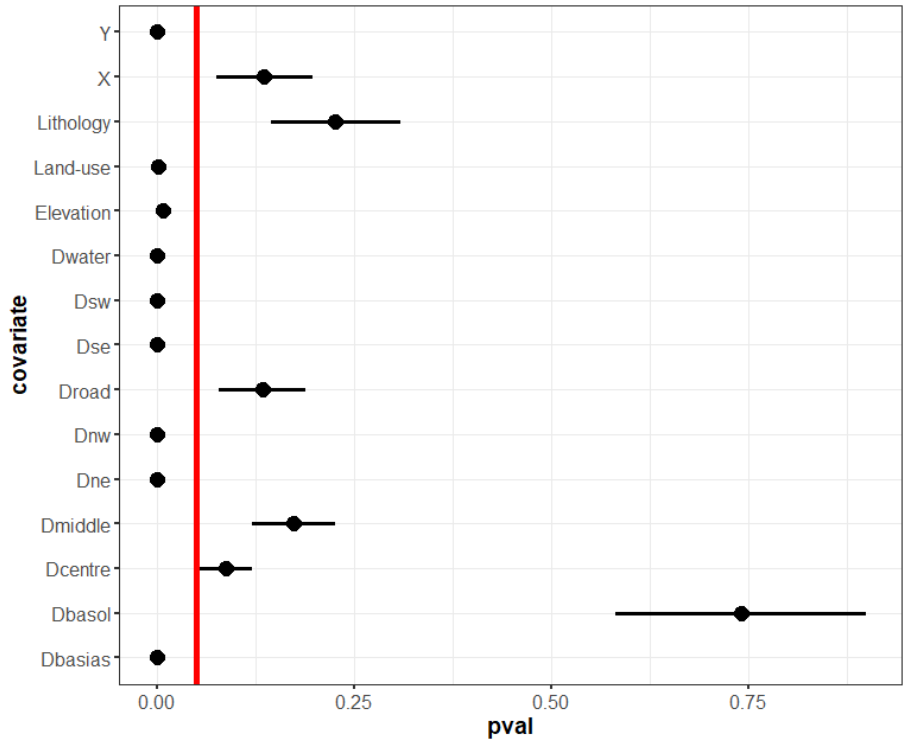
Main methodological concerns

- *My first criticism is related to the first step, that is, the elimination of covariates before model training. This is a common pitfall within machine learning in DSM. The problem is with data leakage which may cause bias, and this occurred when covariates are removed from the entire training data set, and not within for example a cross-validation within each fold. Note that any data preprocessing (e.g., normalisation) dealt with in such a way can lead to data leakage. Data leakage may also cause the model's uncertainty to be lower, and this is then also problematic if interpretative machine learning (IML) methods (like SHAP) are used to analyse the relationships between covariates and the model's uncertainty. In addition, with a model such as random forest, covariate selection is not really required, especially with so few covariates (i.e., 15). I invite the authors to refer to the work such as that of Zhu et al. (2023) for guidance on data preparation so that data leakage is avoided.*

We are grateful to Referee #1 for pointing out the potential problem of data leakage. Now we better underline that the proposed screening analysis is conducted during the cross-validation procedure as recommended by Zhu et al. (2023). Our *HSIC*-based covariate selection is now analysed at each iteration of the cross-validation. Considering the 10-fold cross validation procedure (repeated 25 times), new Figure 10 shows the corresponding p-values. The dots indicate the mean value estimated over the replicates of a 10-fold cross validation (repeated 25

times), and the lower and upper bounds of the error-bar are defined at +/- one standard deviation. . When the dot merges with the error-bar, this indicates that the value of the standard deviation is low.

Covariates with p-values below the 5% significance level are considered influential. This shows that, out of all the cross-validation replicates, nine covariates have a statistically significant influence on hydrocarbon concentration. These covariates are retained in the construction of the RF model.



New Figure 10: Screening analysis showing the p-values of the *HSIC*-based test of independence for the Toulouse case. The dots indicate the mean value estimated over the replicates of a 10-fold cross validation (repeated 25 times), and the lower and upper bounds of the error-bar are defined at +/- one standard deviation. When the dot merges with the error-bar, this indicates that the value of the standard deviation is low. The vertical red line indicates the significance threshold at 5%. When the p-value is below 5%, it indicates that the null hypothesis should be rejected, i.e., the considered covariate has a significant influence on the hydrocarbon concentration, and is retained in the RF construction.

Regarding the usefulness of the screening analysis, we only partly agree with Referee #1's comment, because many real case studies have implemented such approaches to select covariates using either recursive feature elimination or forward feature selection or RF importance measures for cases with both very large number of covariates, such as the study by Poggio et al. (2021), but also with a moderate number of covariates of the order of 10-20 as in our case, such as the study by Meyer et al. (2019) and Dornik et al. (2022).

In our study, we rely on an alternative approach which has proven to be very efficient in the machine learning community (see e.g. Gretton et al., 2005) with applications in multiple domains, e.g. atmospheric pollution (Fellmann et al., 2023); environment (Lambert et al., 2024); geochronology (Herrando-Pérez & Saltré 2024); nuclear safety (Marrel and Chabridon, 2021); deep learning and image analysis (Novello et al., 2022); geothermics (Rohmer et al., 2023). The key advantages in our case are:

(i) *HSIC* measure can capture arbitrary dependence without resorting to some assumptions such as linearity, monotonicity;

(ii) *HSIC* measure can handle random variables potentially of mixed type, continuous or categorical;

(iii) *HSIC* measure avoids the use of RF importance measures, which show some limits as extensively discussed, among others, by Ishwaran (2007), Strobl et al., (2007), Benard et al. (2022). This aspect has also been clearly underlined by Meyer et al. (2019).

The objective of the *HSIC*-based covariate selection is thus two-fold:

- Using a model-agnostic approach that avoids the use of an importance measure that is inherent to the selected ML model;
- Decrease the computational burden of the SHAP approach.

Main correction:

The covariate selection is now conducted in the cross-validation procedure following the recommendation of Zhu et al. (2023). The twofold objective of the *HSIC*-based covariate selection is now better highlighted in Sect. 3.1 'global procedure'. We recognise, however, that it is beyond the scope of this study to compare this method with alternative techniques available in the literature. This is indicated as a perspective.

References

Bénard, C., Da Veiga, S., & Scornet, E. (2022). Mean decrease accuracy for random forests: inconsistency, and a practical solution via the Sobol-MDA. Biometrika, 109(4), 881-900.

Dornik, A., Cheţan, M. A., Drăguţ, L., Dicu, D. D., & Iliuţă, A. (2022). Optimal scaling of predictors for digital mapping of soil properties. Geoderma, 405, 115453.

Fellmann, N., Pasquier, M., Blanchet-Scalliet, C., Helbert, C., Spagnol, A., & Sinoquet, D. (2023). Sensitivity analysis for sets: application to pollutant concentration maps. arXiv preprint arXiv:2311.16795.

Herrando-Pérez, S., & Saltré, F. (2024). Estimating extinction time using radiocarbon dates. Quaternary Geochronology, 79, 101489.

Ishwaran, H. (2007). Variable importance in binary regression trees and forests. Electronic Journal of Statistics, 1:519–537.

Lambert, G., Helbert, C., & Lauvernet, C. (2024). Quantization-based LHS for dependent inputs: application to sensitivity analysis of environmental models. https://ec-lyon.hal.science/hal-04546338/file/Article_QLHS.pdf

Marrel, A., & Chabridon, V. (2021). Statistical developments for target and conditional sensitivity analysis: application on safety studies for nuclear reactor. Reliability Engineering & System Safety, 214, 107711.

Meyer, H., Reudenbach, C., Wöllauer, S., & Nauss, T. (2019). Importance of spatial predictor variable selection in machine learning applications–Moving from data reproduction to spatial prediction. Ecological Modelling, 411, 108815.

Novello, P., Fel, T., & Vigouroux, D. (2022). Making sense of dependence: Efficient black-box explanations using dependence measure. Advances in Neural Information Processing Systems, 35, 4344-4357.

Rohmer, J., Armandine Les Landes, A., Loschetter, A., & Maragna, C. (2023). Fast prediction of aquifer thermal energy storage: a multicyclic metamodelling procedure. Computational Geosciences, 27(2), 223-243.

Strobl, C., Boulesteix, A.-L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics, 8:25.

- *Linking to my previous point. if the goal is to speed up computations, then removing covariates should not be a first choice. In addition, in typical DSM projects the number of covariates is usually more than 100. Therefore, the presented case study, which only has 15 covariates, is not the best choice to showcase the proposed methodology. One could rather perform a sample of grid cells at which Shapley values are estimated. Like for example in the Wadoux et al. (2023) paper. Again, in many DSM projects, maps are sometimes created over millions of grid cells, so the presented case study is not the best one to showcase this methodology. Therefore, to speed up computations with a small data set (like the one in this study), I would rather use a stronger machine to do the calculations*

than to omit potentially important parts of my data. If not possible, then let the computations run for a few days.

We thank Referee #1 for the comments. It is true that the presented case study is not representative of very large scale projects with millions of grid cells and hundreds of covariates. However, we would like to underline that numerous case studies have been found in the literature with number of covariates that are comparable to our case study. Among others, please refer to:

- de Bruin et al. (2022) used a set of 15-20 covariates to predict the organic carbon stock and the above ground biomass;
- Dornik et al. (2022) used 10-15 covariates to predict soil properties in Romania;
- Meyer et al. (2019) used 16 covariates to the classification of Land Use/Land Cover in Germany;
- Fendrich et al. (2024) used 17 covariates to predict arsenic in European topsoils;
- Milà et al. (2022) used 19 WorldClim bioclimatic variables for their synthetic case;
- Wadoux et al. (2023) used 23 covariates for their study.

Regarding the very large number of grid cells, we propose to improve the discussion on this aspect. In the discussion section, we now indicate how the proposed approach could be helpful for these challenging cases. The combination of the grouping and of the screening analysis allow us to decrease the computation cost from 1 day to less than half an hour given approximately 45,000 grid cells. Accounting now for the constraints of global scale studies such as Poggio et al. (2021), a direct SHAP analysis would imply >22 days of calculation, hence requiring a high performance computing architecture. Our approach would here imply <1 day of computation on a single laptop. Moreover, with the growing concern of energy consumption (see e.g., Jay et al., 2024) for scientific computing, we believe of the importance of providing the soil scientists with efficient, energy-saving analytical tools. The other side of the coin is however the introduction of some simplifications that are discussed in the reply to Referee #1's next comment.

Main correction:

While we believe that our actual case is representative in terms of number of covariates of many found in the literature, we recognise that our approach should be better discussed in relation to the challenge of large-scale projects. To this end, we have replaced section 5.2 'Added value of clustering' with a new section 5.2 'Large-scale implementation' to better highlight the challenge of handling hundreds of covariates as well as implementing studies on a national or global scale. This aspect is also highlighted as a perspective of the present study.

References

de Bruin, S., Brus, D. J., Heuvelink, G. B., van Ebbenhorst Tengbergen, T., and Wadoux, A. M. C.: Dealing with clustered samples for assessing map accuracy by cross-validation. Ecological Informatics, 69, 101665, 2022.

Dornik, A., Cheţan, M. A., Drăguţ, L., Dicu, D. D., & Iliuţă, A. (2022). Optimal scaling of predictors for digital mapping of soil properties. Geoderma, 405, 115453.

Fendrich, A. N., Van Eynde, E., Stasinopoulos, D. M., Rigby, R. A., Mezquita, F. Y., & Panagos, P. (2024). Modeling arsenic in European topsoils with a coupled semiparametric (GAMLSS-RF) model for censored data. Environment International, 108544.

Jay, C., Yu, Y., Crawford, I., Archer-Nicholls, S., James, P., Gledson, A., et al.: Prioritize environmental sustainability in use of AI and data science methods. Nature Geoscience, 1-3, 2024.

Milà, C., Mateu, J., Pebesma, E., & Meyer, H. (2022). Nearest neighbour distance matching Leave-One-Out Cross-Validation for map validation. Methods in Ecology and Evolution, 13(6), 1304-1316.

Meyer, H., Reudenbach, C., Wöllauer, S., & Nauss, T. (2019). Importance of spatial predictor variable selection in machine learning applications–Moving from data reproduction to spatial prediction. Ecological Modelling, 411, 108815.

- *The grouping of covariates is a practical way of speeding up computation, but I am afraid it holds no meaning for DSM practitioners. The authors acknowledge this in the discussion, starting at Line 519. Doing inference on machine learning output with IML methods is hard enough. I cannot see how the grouping of covariates could hold much interpretive meaning.*

From Referee #1's comment, we understand that we have falsely conveyed our message about grouping. It should not be understood as a "one-fits-all" method. Instead the grouping should preferably be used to help the interpretation of the Shapley values. This can be done by defining, from the beginning of the analysis, groups of covariates that hold:

- a certain redundancy in terms of information due to the strong dependency between them. This is the solution followed in the original version of our work, but we recognise that it cannot be the only one;
- a meaning for analysists or the end-users. This is the second grouping option by Jullum et al. (2021) based on underlying knowledge/expertise (i.e. grouping covariates that make sense with respect to the problem at hand). As a motivation for this option, the study of Wadoux et al. (2023) is illustrative. In the presentation of their results (Figure 6 of their study), they naturally propose to analyse groups of covariates.

Main correction:

In the revised version of the manuscript, we propose to:

- reformulate our message about grouping by presenting it as an option to facilitate the analysis, and not a mandatory step. The section 3.1 "overall procedure" has been reworked along these lines;
- re-analyse our case study by reworking the groups using the information on dependency and the experts' knowledge.

In our real case, we now analyse four groups of covariates:

- Land-use;
- The elevation.
- The group $D_{basias\text{-}water}$ which includes D_{basias} and D_{water}. Since group reflects the general tendency of industrial sites to locate close to a water supply, the analyse of the joint influence is meaningful;
- The group of geographical coordinates, i.e. D_{ne}, D_{se}, D_{nw}, D_{sw}, and the Y-coordinate. This group of covariates were introduced to improve the predictive capability of the RF model by following the approach by Behrens et al. (2018). Interpreting the respective influence of each of these individual covariates is in practice tricky, and grouping them makes sense in this regard.

- *To sum up, exploring the relationship between covariates and model uncertainty is intriguing and worth exploring. However, the paper's emphasis on reducing computation with (questionable?) methods distracts from the main goal of the paper. That is, I would have liked to see more in-depth analysis of covariates related to SHAP (prediction) vs SHAP (uncertainty). I would also like to have seen more emphasis on: do we expect the same covariates to be related to both, why do we see different covariates in terms of predictions vs uncertainty.*

We agree with Referee #1 that our message on the methods has to be clarified.

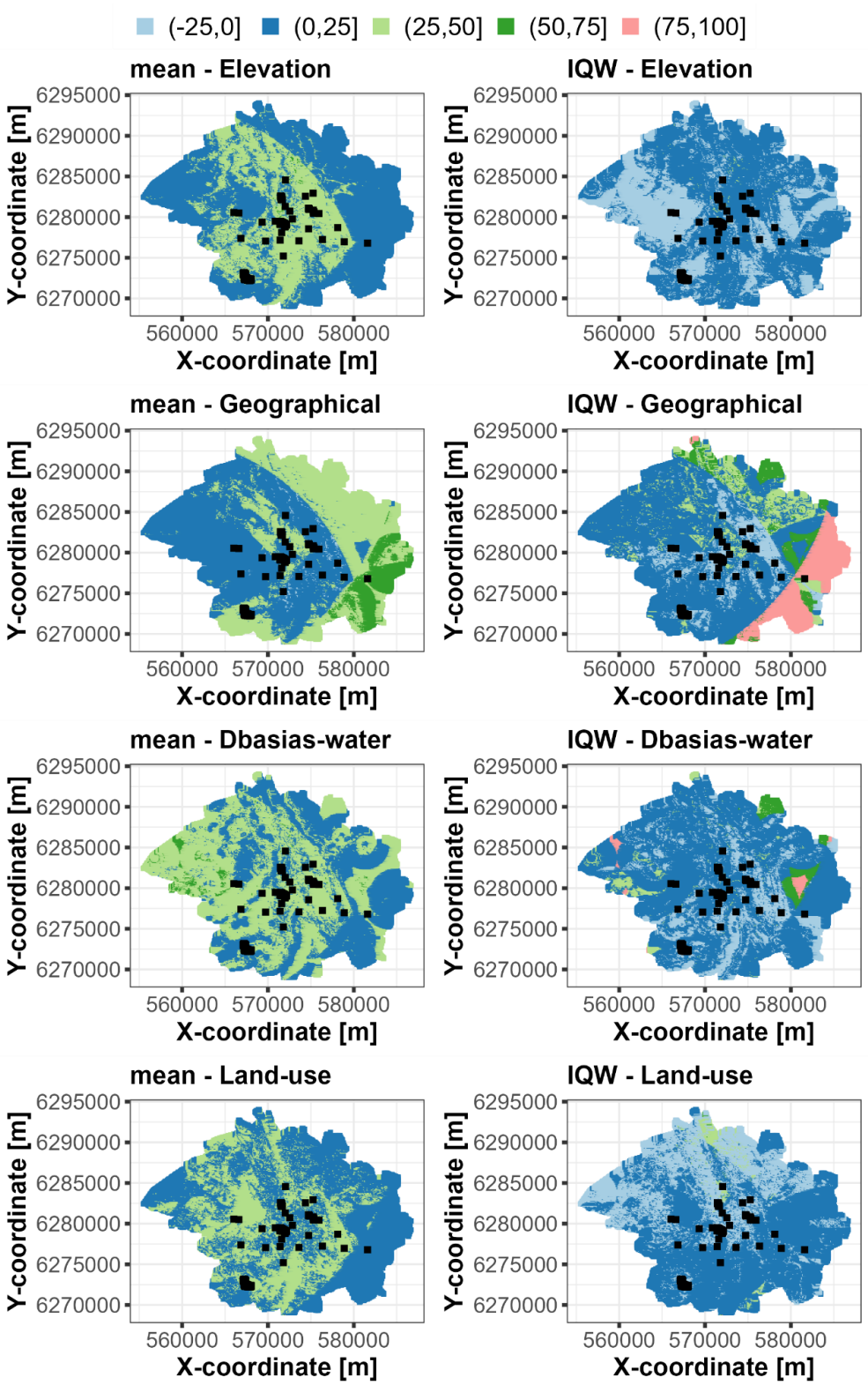
Main correction:

We are now emphasising the advantages of the different stages (elimination of characteristics, grouping).

Particular attention has therefore been paid to

- Improving the implementation of the selection analysis based on the recommendation of Zhu et al (2023);
- Clarifying the definition of groups by combining information on dependency and expert knowledge;
- Improving the presentation of clustering as an option to facilitate the interpretation of Shapley values instead of a single method;
- Further developing the discussion regarding transfer to large-scale projects with millions of grid cells and hundreds of covariates.

We also agree with Referee #1 on the interest of further exploring the link between SHAP (prediction) vs SHAP (uncertainty). To improve this aspect, we propose to define a common level of comparison by normalizing the Shapley values with the same quantity, i.e. the predicted value. New Figure 13 has been updated accordingly (see below). This shows that the scaled Shapley's values are mainly in the range [0, 25%], but with some particular areas where the determining factors for one or other situation (best prediction estimate or uncertainty) are not necessarily the same.

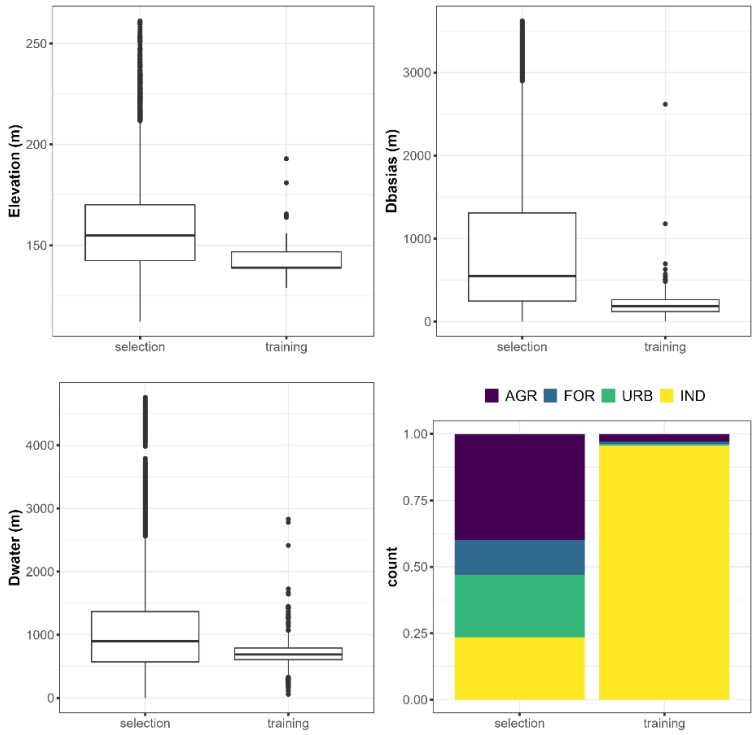


New Figure 13: Scaled Shapley value (in %) for each group of covariates of the Toulouse test case considering the prediction best estimate using the RF conditional mean (left), and the prediction uncertainty using the qRF inter-quartile width *IQW* (right). The black squares indicate the location of the soil samples used for the RF training.

Main correction:

To deepen the analysis of the link between SHAP (prediction) vs SHAP (uncertainty), we propose to investigate in more details three distinct cases which are relevant from the viewpoint of uncertainty management.

- The first case corresponds to locations where at least one group of covariates contribute significantly to the uncertainty, by more than 25% compared with its contribution to the best estimate. This corresponds to <15% of study area;
- The second case is the opposite of the first one, and corresponds to locations where at least one group of covariates contribute significantly to the best estimate, by more than 25% compared with its contribution to the uncertainty. This corresponds to about 66% of the study area;
- Finally, the third situation overlaps with the second case and corresponds to where at least one covariates' group has negative contributions (in light blue in Figure 13, bottom), i.e. where they participate directly in reducing prediction uncertainty. This corresponds to a large proportion of the study area, of about 80%.



New Figure 15: Distribution of covariate values in the form of box plots for locations (termed 'selection') where the corresponding group of covariates contributes significantly to the best estimate, with a Shapley value of the best estimate exceeding that of the uncertainty by more than 25%. These box plots are compared with those for the training data. The bottom right panel compares the proportion of land use categories (AGR: Agriculture, FOR: Forest and Grassland, IND: Industrial and Commercial Economic Activities) for the selection and training dataset.

As recommended by Referee #1, we analyse the analysis of the relationships by examining the distribution of the corresponding covariates. Comparison with the training data set gives us an insight into the reasons for the different situations. New Figure 15 illustrates the second case. It reveals that these locations have elevation values and distances D_{basias} and D_{water} of the same order of magnitude as those in the training dataset. This corresponds to a prediction situation where the RF model is used to predict cases that show similarities to the training dataset. This also means that this prediction situation does not rely too much on the extrapolation capability of the RF model; a situation known to be difficult for this type of ML model (Takoutsing and Heuvelink, 2022). In the areas where land use contributes most to this case, we show (Fig. 15, bottom, right-hand panel) that this is linked to agricultural areas and forests, i.e. areas where there is less chance of finding potentially polluted sites, as shown by the analysis of the training dataset.

It should be noted that, for the sake of brevity of the revised manuscript, some of these analyses are included in the supplementary documents. In particular, the number of figures in the main text has been limited to around ten.

Reference

Takoutsing, B., and Heuvelink, G. B.: Comparing the prediction performance, uncertainty quantification and extrapolation potential of regression kriging and random forest while accounting for soil measurement errors. Geoderma, 428, 116192, 2022.

Some other concerns / suggestions

- *The synthetic case study adds no value to the paper. I suggest removing it as the paper is already a bit long for the topic at hand.*

We only partly agree with Referee 1' comment, because Referee 2 clearly emphasised the value of this synthetic case study in facilitating understanding of the methods. In addition, the data from the Toulouse case study has restricted access. We believe that having a synthetic test case that we can share publicly should facilitate the use and critical analysis of our approach. Therefore we have chosen to keep the synthetic case in the study.

- *Section 3.1 is difficult to follow without the knowledge of HSIC and some of the information in the many cited references. Maybe just restructure the manuscript and include essential methodology.*
- *Random forests are standard and already widely known in DSM. The sections on RF and QRF can be removed, and replaced with brief references to RF and QRF.*

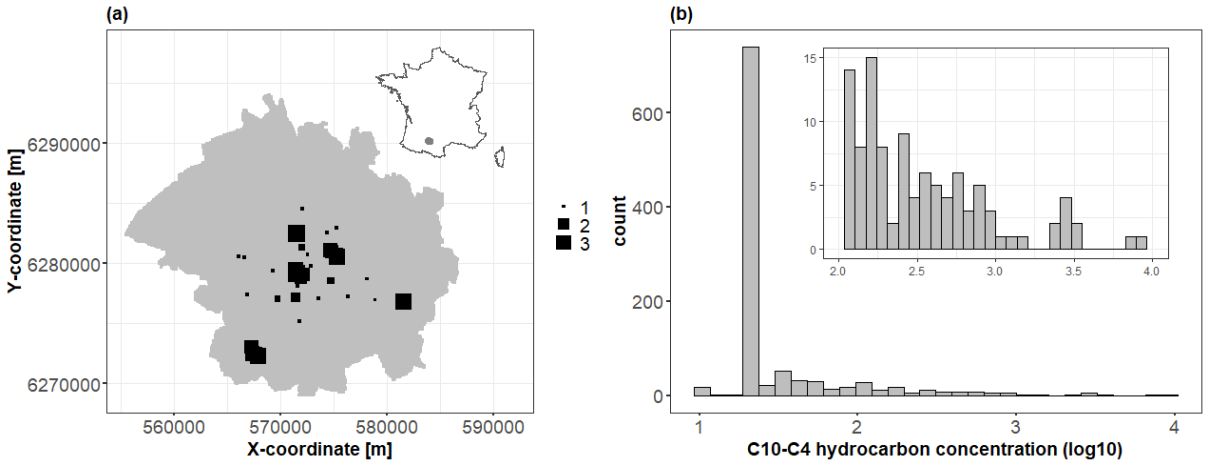
We reply here to both comments. To improve the clarity of the presentation, we have reformulated Sect. 3 by:

- describing the essential details of the methodology in a section Sect. 3.1 "overall procedure";
- describing in full details in section Sect. 3.2 "3.2 Shapley additive explanation" the approach based on Shapley values with additional comments on the computational burden and the benefit of grouping;
- moving the sections on RF models and on HSIC dependence measure in Supplementary Materials.

- *Maps presented in this manuscript are of poor quality and not visually appealing. Captions and legend can also be improved. With Figure 3, show more information. Not everyone is that familiar with this region in France. The histogram is not very clear, especially the long right tail can be enhanced visually.*

The maps have been reworked. In particular the locations of the samples are systematically added to the maps to ease the interpretation of the results with respect to the training dataset. More appealing color scales have been chosen, namely "Set3" and "Paired" from ColorBrewer (https://colorbrewer2.org/); see above an example with new Figure 13. The captions and legends have also been further detailed.

In particular, the presentation of Figure 3 has been improved.



New Figure 3: (a) Spatial location of the 1,043 soil samples (square-like markers) across Toulouse city located in the South-West of France (see the grey dot in the top right inserted map). The size of the squares is proportional to the logarithm (base 10) of the C10-C4 hydrocarbon concentration (expressed in mg/kg). (b) Histogram of the logarithm (base 10) of the C10-C4 hydrocarbon concentration (expressed in mg/kg) with a zoom on the interval 2.0-4.0 (top right inserted panel).

- *General writing of the manuscript is poor. Some examples: The overuse of "etc", too many brackets to give additional information, brief introductions at each section.*

Careful rewriting has been carried out to avoid unnecessary 'etc' and the brackets. The brief introductions have been removed.

- *The mathematical writing can also be improved. For example, are the ahuthors sure that ML model is just y=f(x)? See Line 142.*

The mathematical writing has been cross-checked and the identified problem in line 142 has been corrected as follows: "The mathematical relationship is modelled by a ML model (denoted *f*(.)) sot that $f(\mathbf{x}(s))$ is assumed to resemble $y(s)$ as closely as possible. i.e. $y(s) \approx f(\mathbf{x}(s))$."

- *Figure 6 does not make sense. Why is there an arrow from Step 2 to 4?*

From Referee #1's comment, we have the impression that Figure 6 introduces some confusion. We propose to remove it and to describe in more details the different steps and their interplay in the sub-section Sect. 3.1 "overall procedure".

References:

Wadoux, A., Saby, N., Martin, M. (2023). Shapley values reveal the drivers of soil organic carbon stock prediction. SOIL, 9, 21-38. doi: 10.5194/soil-9-21-2023.

Zhu et al. (2023). Machine Learning in Environmental Research: Common Pitfalls and Best Practices. https://pubs.acs.org/doi/10.1021/acs.est.3c00026.

Orleans,
April 29th, 2024
J. Rohmer[1] on behalf of the co-authors

[1] BRGM, 3 av. C. Guillemin - 45060 Orléans Cedex 2 – France