Recommender 3 review

Answer to the recommender review

Changes in the text

essd-2025-418 "Soil information and soil property maps for the Kurdistan region, Dohuk governorate (Iraq)"

Bellat et al.

Review by Bas Kempen, 23 Jan 2026

A comprehensive paper in a critically under represented geographical area when it comes to soil profiles/ digital soil mapping. Care has been taken with the landscape characterisations including tectonic development and parent material climate and vegetation and geomorphology and soils, as well as maps and photographs to allow the reader to really understand the study area. While the methods are not necessarily 'new' themselves, it is an important application of state of the art methods, the novel part of this study is the study area. The output maps are compared to SoilGrids, a global model, with inputs from WoSIS, the study mentions that WoSIS has low sampling density in this area, highlighting the need for such studies, in addition to the scarcity of other options in the area.

I agree with reviewer 1 that this is a well written and thorough study. This study is a good example of adhering to FAIR and open metadata standards, for not only the data but the methodology, and that is commendable. The methods are comprehensively described, and fit their purpose well. All methodological aspects including the code are not only made available following FAIR principles, but thoroughly documented and explained at https://mathias-bellat.github.io/DSM-Kurdistan/digital-soil-mapping.html#visualisation-and-comparison-with-soilgrid-product, creating a fine example of a fully reproducible study and the input and output data themselves are a much needed addition to more or less non-existent openly available soil data in the area. With this, this manuscript fits well within the scope of ESD.

One of the reviewers mentioned the limited geographic scope of the paper. I do agree that this scope is limited but having a DSM study published for a region like Kurdistan (Iraq) is worthwhile and to me an a welcome addition to the body of literature on this topic. Especially given that the authors make their results as well as data open from which other DSM efforts (e.g. SoilGrids) can profit. This is much appreciated.

Having said this, I do have a few comments and questions regarding the manuscript, particularly concerning the methodologies, which in my view require some further explanation and clarification. I encourage the authors to address these points, after which I would recommend publication of this article in ESSD.

Thanks you for this feedbacks on our study. We are please that our aims, for producing a fully reproducible and open methodology in an under-represented areas, has been acknowledged and highlighted by the recommender. Indeed, we do hope that this study and in particular the results would be used in improving global databases, such has WoSIS, or models as *SoilGrids*.

Main comments

Lines 230-231: Could the authors explain the decision to model each depth layer separately instead of developing one model per property with using the depth as an explanatory variable? Modelling each depth separately is a valid approach, but I would like to understand the reason why the authors took this approach.

Indeed, this choice is motivated by two main arguments:

- First is the question of methodological reproducibility. We based our approach on similar studies, such as the *SoilGrid.2* (Poggio et al., 2021), national soil texture modelling (Varón-Ramírez et al., 2022) – a very well-performed model which inspired part of our methodology - and all these models and maps are using independent depth increment models. Our methodology was also inspired by Malone et al. (2022), who produced independent depth models based on different soil samplings.
- Second is the question of whether or not the soil depth should be used as a predictor. Besides the covariance/collinearity, which would be really high due to the numerous covariates used in both modelling, some scholars have shown that a 3D data-driven approach should be taken with caution (see Ma et al., 2021).

To continue this discussion, we considered setting a "mask" for soil depths below 5 cm. However, the purpose of this study, and of ESSD to some extent, is provide more **raw** data that can be interpreted by any researcher regarding their own discipline. A geochemist or people working with remote sensing might find some interest in having, even if low, a partial soil covering over the whole study area. While researchers working on pedogenesis of deeper soils will likely exclude all soils with depths lower than 5–15 cm.

Lines 246-247: I do not understand the rationale for combining data splitting (80/20) with (repeated) cross validation. Cross-validation already produces an independent prediction for each data point, from which accuracy metrics can be calculated. What was the motivation for embedding CV within a data-splitting framework? And how does this then work? If CV was performed on the 80% training subset, this would give CV-predictions only for these points. I wonder how predictions for the 20% test set were obtained. Which trained model was used to predict at the points in the test set? In case of normal data splitting this would be the model trained on the training dataset. However, by running CV on the training dataset there is no single trained model but multiple (here 10) fold-specific models.

Thank you for this comment. You raise an important point on the preparation and training of models. We do think there is often a misunderstanding between **validation** and **test** data. The validation set serves as an independent set during training to tune the model's parameters. The test values serve only as a "control" of the model based on the evaluation process at the end. This set is then fully independent of any prior training and better reflects the model's adaptability and **interoperability.** The introductory book by Alpaydin provides a well-explained analogy of the **train/validation/test sets** (Alpaydin, 2014, p. 40).

Regarding the selection from this CV, we selected the best model and evaluated it on the test set, reporting the **test evaluation metrics** in the table. We used the `*train*` function from *caret* and the basic `*predict*` function from stats R to predict the test set as :

"predict(Evaluation$Models[[i]][[j]], X_test)"

The final will be the training being the `*Evaluation$Models[[i]][[j]]$finalModel*` which is a model based on the training with the best hyperparameters collected from the 10 folds of each 3 repetitions ending in a total of **30** models trained with the tuning grid (depending on each type of model used). For example, for the **Cubist** model, we used the following tuning grid "*expand.grid(committees = c(1, 5, 10, 15, 20), neighbors = c(0, 1, 2, 3, 5, 7, 9))*". In total, for one **Cubist** training on a variable, we have 35 (hyperparameters) x 30 (resampling) = **1050** independent models.

The `*train*` function then reselects the best combination as explained in the ***Caret*** manual: "*The combination with the optimal resampling statistic is chosen as the final model and the entire training set is used to fit a final model.*".

We did add a sentence to detail the split strategy with clarification on **validation/test** sets:

"For each soil depth, we used a two-step evaluation. First, we set aside 20% of the data as an independent test and did not use it for any model decisions. Then, we used the remaining 80% as the training set and ran repeated k-fold cross-validation to tune the model and choose the final settings. After the final settings were chosen, we trained the final chosen model on the full 80% training set and used this single model to make predictions for the held-out 20% test set."

In addition, the manuscript states that CV was repeated three times? While repetition may improve robustness when data are limited or when using a small number of folds, with 10-fold CV I would expect only minimal variation between the repeats?

The point of repeating the CV is not to improve the model but to reduce bias sampling. As our training sample sizes are quite small (98, 90, 86, 87, and 74, respectively, due to the test set split), and we are handling spatial data that is often prone to overfitting (Brus, 2022), we need to reduce the bias introduced by our sampling strategy.

Furthermore, the RMSE values varied slightly across repetitions, suggesting that a single k-fold CV could have yielded a biased estimate, especially given the spatial nature of the data. The repeated CV provides a more stable and reliable assessment. A recent study by Lumumba et al. (2024) found lower variance and greater stability of cross-CV compared with standard CV.

We therefore assessed each model's variability to determine whether any fold showed massive overfitting. **None was witnessed (< 10% variance in the RMSE)**, and an extra line of code was added to a new table for each depth, "Repetition_variance.txt," in the supplementary files.

We did add an extra sentence to clarify our use of the three-time repetition CV:

"This resampling strategy allows us to avoid potential overfitting due to the small size of our training data set (< 100) and the spatial nature of our data."

Overall, the validation approach seems a bit overcomplicated. The authors may well have had sound reasons for adopting this approach, but in case the rationale and precise implementation need to be explained more clearly.

We already answered part of the question in the above comments. However, to complete our answer, we do not think the resampling process chosen is inadequate. We are facing, as many DSM studies have, two main issues:

- The bias in our samples due to their spatial nature, inducing either a high number of repetitions - our chosen approach – or a spatial cross-validation such as the *BlockCV* package (Valavi et al., 2019) or another type of spatial sampling (Schwarz et al., 2019).
- The size of our data is rather limited, less than 100 samples for the training phase. The number of repetitions has to be consistent to reduce variance and limit overfitting. An alternative approach would have been to use leave-one-out (LOOCV) sampling.

To address both issues, we adopted a 3-fold repeated 10-fold CV. As it might be more time-consuming than a "classical" CV (Lumumba et al. 2024), it yields less variance than a simple CV and is less demanding in terms of computer resources than a spatial LOOCV.

Overall, if the three repetitions do not improve the accuracy or sensitivity of the models, they do help reduce the potential bias from overfitting.

Lines 268-273: The description of the ensemble modelling approach is unclear to me and would benefit from additional detail. Specifically, it is not clear which 'conditions' (l. 269) are being referred to, what criteria were used to select 'the best one', and how the individual model predictions were combined in the ensemble? While relevant literature is cited, I believe that a few additional lines better outlining the implementation of the ensemble approach would improve clarity to the reader.

Sorry for the unclearness of our sentence. You are right, the term "conditions" is totally misused here. We used a stacked regression. We did clarify the meaning of the ensemble model and the "meta-model" used here.

"Finally, an ensemble model is created by stacking the five previously trained models. A random forest serves as the "meta-learner", which takes the predictions of the base models as input and learns how to optimally combine them using the caretStack function."

Other comments

Line 10: The summation signs should be removed I believe.

We did remove the signs as in line 14 of the abstract.

Line 10-11: Reference is made to 'local models' (compared to the regional models the authors developed and the SoilGrids global model). It is unclear to me where the 'local models' refer to and what the basis is for the claim of the authors.

Thank you for pointing out this mistake. We wanted to refer to 'regional' models, which are intended for use at finer scales (e.g. regions, cities or towns) than national or worldwide models. We did change the term "local" to "regional". We do not differentiate between our models and the ones from Yousif et al. (2023) in term of scale.

Line 14: I believe the minus sign in the superscript should be removed, assuming the RMSE values are reported for the transformed depth data. After applying the square-root transformation, the depth unit becomes cm^0.5. The unit of the MSE metric would then be cm, and taking the square root to obtain the RMSE would again give value in cm^0.5. Table 5 also reports the RMSE unit in cm^0.5. I assume the unit of the MAE (Table 5, l. 337) is also cm^0.5? ~~The '-0.5' superscript should be removed I believe (same for line 337).~~

Thank you for noting this mistake. However, we recomputed the soil depth with a fully reproducible workflow (adding python GEE API), during the update we changed the code to produce directly transformed metrics. We therefore, removed all the reference to transformation ($cm^{0.5}$) in the text.

Line 148: Reference is made to WRB 2006. Can the authors confirm if this was also given that there are more recent versions of the WRB? The latest from 2022 I believe.

Comments from recommender 1 were pointing out to this mistake. We did change all the references according to WRB 2022, which did not yield major difference for aridic and semi-aridic soils.

Line 178: "potential soil layer" - inconsistent naming – previous paragraph is "potential soil properties"

Thank you for point out this mistake, a change was done.

Line 185: What are the 'layers' here? Are these the soil horizons or are these the layers for which samples were collected (l. 186)? Please clarify.

You are entirely right this induced to a misunderstood. We do mean "horizons" in here.

We changed for "soil horizons' depths"

Line 187: How is 'topsoil' defined here? Is this the 0-10 layer? Explain a bit more how the ring samples were taken. E.g. where was the ring sample was taken in the topsoil layer: from the top, in the middle of the layer …?

Sorry for the lack of clarity in this element. We sampled the top of the soil after removing all vegetation and loose material. Related to the height of the ring (c. 5 – 7 cm), only the upper part of the top soil, 0 – 10 cm, was then sampled.

We rephrase the sentence to clarify:

"After removing the surface litter and loose sand, the sampling ring was used on the 0 –10 cm soil layer."

Lines 236-237: References seem to be incomplete. Only years are mentioned (2021,2022)

Sorry for this error the new version do not have this typo. The related references are (Poggio et al., 2021; Varon-Ramirez et al., 2022).

Line 243: Could expand on why RFE was not restrictive enough/ restrictive enough for what?

The results of the RFE did provide optimal tuning for a larger number of covariates (*c.* > 20) compared to the relatively low number for the Boruta (< 10). Therefore, the computing time needed for modelling would have been greater for limited results.

Details of the RFE results can be found in the supplementary material (7 – DSM/export/RFE), including a detailed graph of the optimal RFE selection and the exported variables selected by RFE.

Line 372: Why were SoilGrids extremes removed?

This question is most welcome. The *SoilGrids* relies on land use imagery for masking some areas (rivers, lakes, oceans or cities). Due to the coarse resolution of the models (250 m) these areas and their neighbouring pixels are often highly outliers. To overcome this effect two option were possible, either using the same mask (which is not provided) or smoothing these outliers. We choose the second

option mainly for practical issues and as the number of pixel presenting outliers were not highly consistent (*c.* < 5%).

Line 401: Limitations are addressed in the paper – I am not sure if the point density, even though notably higher than WoSIS, still warrants mapping at 25m, the reference to Hazelton and Murphy (referencing cartographic scales) seems a bit of a jump. Instead of referencing Hazelton and Murphy I would rather compare to other regional DSM studies.

You are entirely right this was also an underlying comment for the recommender 1. This comparison with "traditional mapping" density from Hazelton and Murphy (2016) is not relevant. We removed the entire sentence.

The Conclusion section reads like an abstract. Reviewer 1 already commented on this and based on that, the cauthors revised the text and I believe with that revision this comment is addressed sufficiently.

Thank you for having reading the revised version of the conclusion. Indeed we changed in depth this conclusion regarding recommender 1 comments. We provide here the text if ever it is not easily readable on the comment version from ESSD.

l. "We developed a complete workflow for digital soil mapping at a regional scale in the Dohuk Directorate of the Kurdistan Region of Iraq, combining cLHS-driven sampling, MIR-based soil property prediction, and several machine-learning models to produce 50 soil property maps at 25 m resolution, as well as regional soil depth and soil class maps. Compared with SoilGrids 2.0 and earlier local products, our models offer more locally relevant predictions and improved spatial detail, while also covering a broader set of soil properties and depth increments than previous regional studies. The soil class map further aligns with current WRB standards and benefits from greater observational density than earlier exploratory works.

Beyond these technical achievements, the study highlights the importance of integrating local measurements with models tailored to regional environmental gradients. Global products provide consistent baselines, but they cannot fully capture the geomorphological and topographic contrasts that drive soil variability at fine scales. The superior performance of our regional models demonstrates the complementarity between global and local approaches: global datasets remain essential for broad-scale comparisons, whereas locally calibrated workflows are crucial for operational land management, agricultural planning, and resource assessments.

The proposed workflow is fully transferable to other regions of similar size (~2,000 km²). Areas with comparable environmental conditions—such as western Iran, northern Syria, or parts of the Mediterranean basin—represent suitable candidates for direct methodological transposition. In addition, the time investment required for the entire process, from sampling design to final DSM production, is relatively modest: in our case, approximately one year (235 person-days). The datasets produced in this study also offer broader reusability, for example, as calibration material for MIR/FTIR spectral libraries (Safanelli et al., 2025; Viscarra Rossel et al., 2016) or as part of a regional or global soil profile archive database (Lachmuth et al., 2025).

A further contribution of this work is the provision of the first regional FAIR-compliant soil dataset (Crystal-Ornelas et al., 2022). Soil science research remains highly geographically imbalanced, with five countries producing more than 80% of global output (Cherubin et al, 2025). Southwestern Asia is sparsely represented, except for Iran, and no soil profiles from the Kurdistan Region of Iraq appear in the WoSIS database (Batjes et al., 2020) used by SoilGrids 2.0. (Poggio et al., 2021). Such data gaps reinforce global inequalities in environmental knowledge (Allik et al., 2020; Sonnenwald 2007) and limit the capacity of data-poor regions to benefit from international modelling initiatives. By openly sharing our dataset and workflow, we help reduce this imbalance and contribute to greater transparency and reproducibility in regional soil information systems.

Overall, this project demonstrates that high-resolution, locally informed digital soil mapping is both feasible and highly effective in data-poor regions. The workflow presented here substantially advances soil knowledge in Dohuk and provides a generalisable and reproducible model for improving soil information in other regions facing similar environmental and data constraints."

Figure 2: What is a negative site? – soil samples

Sorry for not explaining this element. It refers to the line 275 – 276: "The soil depth was measured from 0 to 100 cm on the 122 sampling sites; we added 25 zero values from remote sensing imagery observation on bare rock points.". We did added the mention of **negative site**

" sampling sites; we added 25 zero values (negative sites) from"

Technical corrections

- *Line 21: influence local ecosystems -> influence on local ecosystems*
- *Line 28: includes -> include*
- *Line 33: gives information on its ability to fit or not for agricultural purposes, but also to better understand -> gives information on its ability to fit agricultural purposes and helps to better understand*
- *Line 35: Governate -> governate (not capitalised anywhere else in the paper)*
- *Line 70: cluster -> conditional*
- *Line 75 & 265: Hengl and Robert -> Hengl and MacMillan*
- *Line 74: a raw -> one raw*
- *Line 129: climax -> climate*
- *Line 255: McBradney -> McBratney*
- *Line 229: remove 'part'*
- *Line 234: a additive -> an additive*
- *Line 247: state of the art the art -> state of the art models*
- *Line 270: approached -> approach*
- *Line 346: bakns -> banks*

- *Line 388: remove ‘consistent’*
- *Line 401: Hazelton and Murphy pg5 -> pg4*
- *Line 410: LU/C -> land use/cover (it is not used previously)*
- *Line 425: profiles depth measurement -> profile depth measurements*
- *Line 441: world -> global*
- *Line 443: shallower resolution -> higher resolution*
- *Table 3: Modis brightness index in the wrong column*
- *Appendix B: units column could be named differently. It contains a mixture of units, formulas, ranges.*

We deeply thank the recommender for all of these comment. We did make all the changes needed. As for the Appendix B we did change the name of the column to **“measure”.**

References:

Alpaydin, E. (2014). Introduction to machine learning (Third edition). The MIT Press.

Brus, D. J. (2022). *Spatial Sampling with R* (1re éd.). Chapman and Hall/CRC. https://doi.org/10.1201/9781003258940

Ma, Y., Minasny, B., McBratney, A., Poggio, L., & Fajardo, M. (2021). Predicting soil properties in 3D : Should depth be a covariate? Geoderma, 383, 114794. https://doi.org/10.1016/j.geoderma.2020.114794

Malone, B., Stockmann, U., Glover, M., McLachlan, G., Engelhardt, S., & Tuomi, S. (2022). Digital soil survey and mapping underpinning inherent and dynamic soil attribute condition assessments. Soil Security, 6, 100048. https://doi.org/10.1016/j.soisec.2022.100048

Poggio, L., de Sousa, L. M., Batjes, N. H., Heuvelink, G. B. M., Kempen, B., Ribeiro, E., & Rossiter, D. (2021). SoilGrids 2.0 : Producing soil information for the globe with quantified spatial uncertainty. *SOIL*, *7*(1), 217-240. https://doi.org/10.5194/soil-7-217-2021

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2019). Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. Ecological Modelling, 406, 109-120. https://doi.org/10.1016/j.ecolmodel.2019.06.002

Varón-Ramírez, V. M., Araujo-Carrillo, G. A., & Guevara Santamaría, M. A. (2022). Colombian soil texture : Building a spatial ensemble model. *Earth System Science Data*, *14*(10), 4719-4741. https://doi.org/10.5194/essd-14-4719-2022

Yousif, B. S., Mustafa, Y. T., & Fayyadh, M. A. (2023). Digital mapping of soil-texture classes in Batifa, Kurdistan Region of Iraq, using machine-learning models. Earth Science Informatics, 16(2), 1687-1700. https://doi.org/10.1007/s12145-023-01005-8

Citation: https://doi.org/10.5194/essd-2025-418-RC3