Referee's review

Response to the recommender's assessment.

Changes in the text

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

Bellat et al.

We would like to thank all the referees for their work in discussing our manuscript. We also would also like to thanks the editors as they allowed us to push this pre-print into the review stage.

We have already answered all the referees in detail with separate comments. However, we provide here an overview here of all the major changes applied to our manuscript and also some revisions to previous answers to the referees, which may have changed due to the adaptation of a new workflow in line with referee #4's comments.

Major changes:

- Adoption of a 30 m pixel grid instead of the 25 m pixel grid. We also added extra steps in the pre-process to extract all the covariates in a fully reproducible way from GEE *via* a python API. This was not based on the referee's comment but our own decision toward more "FAIRness" for this study.
- Clarifications on the horizon, topsoil and other soil definitions.
- Only a single QRF model was used for the prediction models with a 10 CV k-fold repeated 3 times strategy. This replaced the diverse algorithms used before, and the split strategy was added on top of the CV.
- Adding more detail on the equations, notably the RUSLE model equations.
- Reduction to fewer metrics (ME, RMSE, R2, RIPQ, PCIP).
- Adding a results part to the laboratory measurements.
- Adapted the section in the results to the interpretation of the evaluation process from the FTIR and DSM predictions.
- Restructuring the conclusion, considering referee #1's comments.

Specific changes:

We listed all the previous answers to referees #1 and #3 here, which have been amended. They have also been provided individually in the discussion forum.

Referee 1:

L233 "We performed a standardisation of the predicted values of the texture on 100 % with TT.normalise.sum function (Moeys et al., 2024) and a additive-log ratio transformation (Aitchison, 1986) with the alr function (Tsagris et al., 2025)." This is not clear. Was the normalization following

the MIR inference/wet lab measurements? And then were the alr variables used in the mapping, followed by back-transformation (as is done in SoilGrids v2.0)?

We removed this part on the scaling of the covariates of the DSM due to the adaptation of our methodology (cf. referee #4 comment).

\S4.3 Another interesting comparison with SG2 would be the prediction ranges. SG2 likely smooths more than this study, see Table 7 where the Q1-Q3 range is always much narrower. This can be brought out in the text -- the interesting discussion is about global vs. local models. The SG2 maps are much more uniform than the maps from this study.

Changes have been made in this section. *SoilGrid 2.0* is now compared to our model without the SD "harmonisation" – removal of the value outside of the SD – and the section has been re-written in consequence of the newly computed methods.

"To assess model performance, we compared our results with the global SoilGrids 2.0 product (Poggio et al., 2021), focusing on pH, OC, Nt, and texture attributes (Table 7), for three generalised depth intervals (0-30 cm, 30-60/70 cm, and 60/70-100 cm). We scaled our predictions to match the 250 x 250 m resolution of SoilGrids 2.0, with a bilinear method from the terra package (Hijmans et al., 2025).

Our models predicted higher values of OC and Nt, with respective increases of $\approx$ 1000%and 300%over those from SoilGrids 2.0. Predicted sand values were also higher (by $\approx$ 25%), while clay values were slightly lower (by $\approx$ 15%). The differences in silt and pH values were negligible, with a 3% higher value for silt and 5% lower for pH in our predictions compared to SoilGrids 2.0.

The standard deviation of the SoilGrids 2.0 product is smaller, except for the pH, than for our prediction models (Table 7). This shows a narrower distribution of values, likely 455 due to the wide range of input data used for the SoilGrids 2.0. The diversity of soil types and input data at the global scale makes the SoilGrids 2.0 model respond relatively homogeneously at the regional scale. Furthermore, the SoilGrids 2.0 product shows a more skewed distribution for OC and Nt, with a higher concentration of values near the lower end of the distribution, which is consistent with the known underestimation of these properties in global models (Shi et al., 2025).

We also compared the evaluation metrics of our predicted values with those obtained from SoilGrids 2.0 on an independent data set (Appendix F). Before training the models on a full data set, we splited the data retaining 20% for test and 80% for training. The models were trained in similar conditions as our main prediction models (cf. 2.4.2), before evaluation were computed on the independant test set. Overall, our models outperformed SoilGrids 2.0 across all evaluation metrics for pH, Nt, OC, sand, silt, and clay at all depth intervals. The only exception was the sand QRF model RMSE score at the 10 – 30 cm depth interval, which was slightly higher for the SoilGrids 2.0 model at the 15-30cm interval."

Referee 3:

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.

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?

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 amended our methodological approach for the DSM computation to apply changes according to referees #3 and #4 by removing the data split part and relying only on the 10 k-fold CV (repeated 3 times).

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.

Due to the change in our methodology this comment no longer applies.

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

We added a supplementary line on this topic.

"We also performed a recursive feature elimination (RFE; Guyon et al. 2002) on the covariates with the caret package (Kuhn, 2019). The results were more conservative with the number of covariates selected (> 60 for each variables), longer in time computing capacities (800%), and provided lower accuracy scores compared to Boruta selection, for the tested 0-10 cm depth increment."

Line 372: Why were SoilGrids extremes removed?

The extremes of the SoilGrids were not removed in the new version, only min. values present "outliers" but they do not influence the global results (Table 7, Appendix F).