*Referee 5 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 Anonymous Referee #5, 02 Feb 2026

*Soil characteristic data are crucial for both soil environment research and model parameter input, especially when large amounts of monitoring data are involved. The authors not only collected extensive samples but also conducted regional simulations, achieving a shift from point-based to area-based data collection. However, some statements in the manuscript do not align with the provided tables/figures, and careful revision by the authors is recommended.*

We deeply thanks the Referee 5 for his time and his appreciation of our work. This comment underlined perfectly the aim of our work which was to performed a quantitative soil characteristic for an area base collection in an underexplored part of the globe.

*The abbreviation for organic carbon is generally OC.*

Thank you for this comment we did replace all the mention of “Corg” by “OC”.

*Are the two sides of the soil samples module in Figure 2 conducted simultaneously, yielding different output result maps respectively? For the rectangle at the bottom of the figure and the text below it, which part of the upper diagram do they correspond to? This is not clear.*

We are sorry if Figure 2 was not clear enough for the reader. We made a small update after changing the methodology for referee 4's comments. The first part of the soil sample refers to the conditioned Latin hypercube sampling selected sites, while the second corresponds to the sites with their z dimension from the different collected depths. Indeed, the cLHS provided site locations, while after sampling, we had different sample counts per site depending on how many depth increments were collected. The total number of sites and sampling used is summed up in Table 1. As for the rectangle below the figure, they were referring to the different types of operations conducted during the workflow, link to the colours in the upper part of the figure. We removed the rectangle as they lead to confusion.

*Does Appendix 1 refer to Table A1 and Table E1 in Appendix B? The order of the appendices seems a bit confusing. Are the details of the method in the supplementary material? Which supplementary materials are you referring to? I can't seem to find the detailed discussion of the method.*

Sorry for the Appendix labelling, as it is our first submission *via* the ESSD LaTEX template we did not mastered all the subtilities of the process. As for the supplementary materials there were available for downloading directly from the pre-print webpage. All the details of our methods, especially for the sampling campaigns, are furnished inside the supplementary material. We provided a full table of each sites location and description we attached photos. For the laboratory

analysis we provided all the details of our measurement with additional plots. Similar for the spectra observations were all data transformations are described with the *R* code corresponding to the operation.

*2.1 Study area and Figure 3: It would be best to label abbreviations for the two parts on the Figure. Additionally, the blue dashed line representing the 2022 sampling area is just a line segment, not an enclosed region.*

I fear we do not understand the first comment on the labelling of the abbreviations in Figure 3, as there are none. Regarding the 2022 campaign sampling area, we rearranged the lines in the figure as also suggested by Referee #1.

*2.1.2 Climate and vegetation: Can you provide a distribution map of the vegetation types for the region?*

Sadly, the Zohary (1973) map is at a regional scale, with poor detail regarding the different compositions in our study area. Nonetheless, Guest and Al-Rawi (1966) provided a more detailed map on p.66 of their book. We did provide an additional Appendix (Appendix B) reproducing this map.

*I cannot find Appendix 2.*

Sorry once again this is a labelling issue, the reference to appendix 2 line 137 refers to the geomorphological map (Appendix C).

*Line 175: I cannot find Annexe 1*

Similar to the previous comment, we are deeply sorry for the mistake in the labelling. This refers to the Appendix A, the table of all used covariates.

*2.3.2: Equation 1 should be provided at the end of this section, with an explanation of what each letter in the Equation represents.*

Thank you for this comment. This joins the referee 1's comments on the lack of equation explanation and referee 4's comments on the location of the figures, table, and equation. We moved the equation to the end of section 2.3.2 and added explanations for each letter.

*Where's the Appendix 3?*

Sorry for the mislabelling this was a reference to the Appendix C.1 (now E.1) with the Boruta selection figures.

*The meaning of each letter in Equations 2-9 needs to be provided in detail.*

Thank you for this comment. This joins referee 1's similar comment. We provided all the letter explanations for the equation. Additionally, we provided the equations for the RUSLE model in the Appendix D.

*Where is Appendix 2? What are the 25 environmental factors included?*

Once again, this is a mislabelling. Appendix 2 (line 278) refers to Appendix A, which lists the selected covariates. The column "Used for" gives the information on which part of the workflow the covariates were used. In the case of the soil depth, the covariates used were: Landsat 5 (Green, Blue, Red, NIR, NDVI, NDWI, LST (for four periods), geology, geomorphology, landuses, PET,

Precipitation, Solar radiation, Wind speed, Aspect, DEM, General curvature, MrRTF, MrVBF, Plan curvature, Profile curvature and TPI. This is provided in Appendix A.1

*Table 4: What's mean for the Q1 and Q3?*

This label where for the 1st quantile and 3rd quantile. We changed it for full name of the quantile.

*2.4 Models and pre-process: This section should be described in conjunction with Figure 2. Currently, the method seems somewhat disjointed between parts. Based on Figure 2, it would be better to provide an overall description of how each step unfolds and operates.*

Thank you for this comment. Regarding the comments of referee 4 with restructured in depth the section 2.4. We hope this new version is clearer and more jointed between the parts.

Here is the new version of the section

"We based our soil property model on the soil formation factors of the scorpan equation (Equation 1) developed by McBratney 240 et al. (2003). We included 85 covariates (Table 2 and Appendix A). The remote sensing variables were accessed through an API of Google Earth Engine (https://earthengine.google.com) on Python, via the ee library (LLC, 2025), and the different indices computed in R with the terra (Hijmans et al., 2025) and raster package (Hijmans, 2010). The terrain variables were computed on SAGA GIS 9.3.1 (Conrad et al., 2015) based on a filled and filtered DEM from GLO-30 ESA and Airbus (2022). All the computation was realised under R 4.4.0 environment (Team et al., 2024). We included only the 2022 and 2023 samples to produce the DSM for two reasons. First, these campaigns followed a cLHS sampling strategy (cf. 2.2.1), whereas the 2017-2018 campaigns did not; including them would have reduced the consistency of the sampling design. Second, the 2017-2018 samples lacked depth information and were primarily limited to topsoil. To ensure data comparability, these earlier samples were therefore excluded, resulting in a dataset of 531 samples from 122 sites (Table 1). These samples were included in the DSM as input with: 122 samples for the 0-10 cm depth, 111 for the 10-30 cm increment, 108 for the 30-50 cm depth, 98 for the 50-70 cm increment and 92 for the 70-100 cm depth. We divided the mapping of each variable for each soil depth increment, resulting in 45 models and 50 maps in total (the three texture variables only include two alr models). We performed a standardisation of the predicted textures values from the Cubist model, with TT.normalise.sum function (Moeys et al., 2024) and an additive-log ratio transformation (Aitchison, 1986) with the alr function (Tsagris et al., 2025). This transformation preserved the spatial information of the prediction with a repartition close to a normal distribution (Liu et al., 2022). Digital soil mapping have adapted this additive-log ratio on the texture with success, alr_sand = ln(sand/clay) and alr_silt = ln(silt/clay) (Poggio et al., 2021; Varón-Ramírez et al., 2022). Once the models were performed, the additive-log ratio was reversed into the three texture with the alrInv function, before being evaluated.

During the pre-processing, we performed a feature selection with the Boruta package (Kursa and Rudnicki, 2010). Using a random forest-based model, Boruta validated or rejected the selection of variables regarding their influence on the inputs (Appendix D). This method improves model accuracy and reduces overfitting results (Kursa and Rudnicki, 2010), and its efficiency has been proven for digital soil mapping (Taghizadeh-Mehrjardi et al., 2020; Suleymanov et al., 2024; Bouslihim et al., 2024). 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 265 depth increment.

For each soil depth, we used a 10 k-fold cross-validation repeated 3 times to tune the model and choose the final settings. This resampling strategy allows us to avoid potential overfitting due to the small size of our training data set (< 100). We trained the models on a quantile regression forest model (QRF), which has shown good performance for digital soil mapping (Varón- Ramírez et al., 2022; Shi et al., 2025). The model was implemented with the caret (Kuhn, 2019), and quantregForest (Meinshausen and Michel, 2020) packages. We tuned the mtry hyperparameters (sequence of 1 to the number of covariates, by steps of one), which corresponds to the number of covariates randomly sampled as candidates at each split, and the minimum node size hyperparameter nodesize (sequence of 5 to 31 by one 5, Shi et al. 2025), which defines the minimum number of samples required to be at a leaf node. The number of trees was set as default at 500 (Liu et al., 2022).

Regression trees use a tree-based structure, splitting the data into different nodes. In the end, the model evaluates the leaves and selects those with the best performance. Their specificity in regression is to predict continuous values at the terminal nodes rather than classes, unlike classification trees. Random forests build upon this principle by combining many regression trees grown on bootstrapped samples of the data, which improves prediction performance and stability. Based on this random forest framework, the quantile regression forest (QRF) (Breiman, 2001), tracks each sample's value at each node, providing a conditional response distribution. This allows the model to produce prediction intervals and to assess accuracy through quantiles (Vaysse and Lagacherie, 2017)."

*3 Result: In the results section, data analysis of the soil characteristics from the collected samples should also be presented. In many cases, actual measured values are more important. Additionally, the data in Table 2 should be expressed in the form of mean ± SD. Additionally, a distribution figure of the measured values could be provided.*

You highlighted an important point on the data quality. We entirely agree with you on the necessity of providing the soil measurement data. Therefore, we added a new subsection in the results section on the measurement results, including a new distribution plot of the measured values (**Figure 7**). However, by lack of space, we provided in the manuscript only the version with a single distribution plot for each variable, therefore mixing the different depths. In the supplementary materials, you will find a plot divided between each soil increment (in the form of boxplots).

We did not discussed the raw results as no regional or national data set exists it is difficult to compare it with something similar.

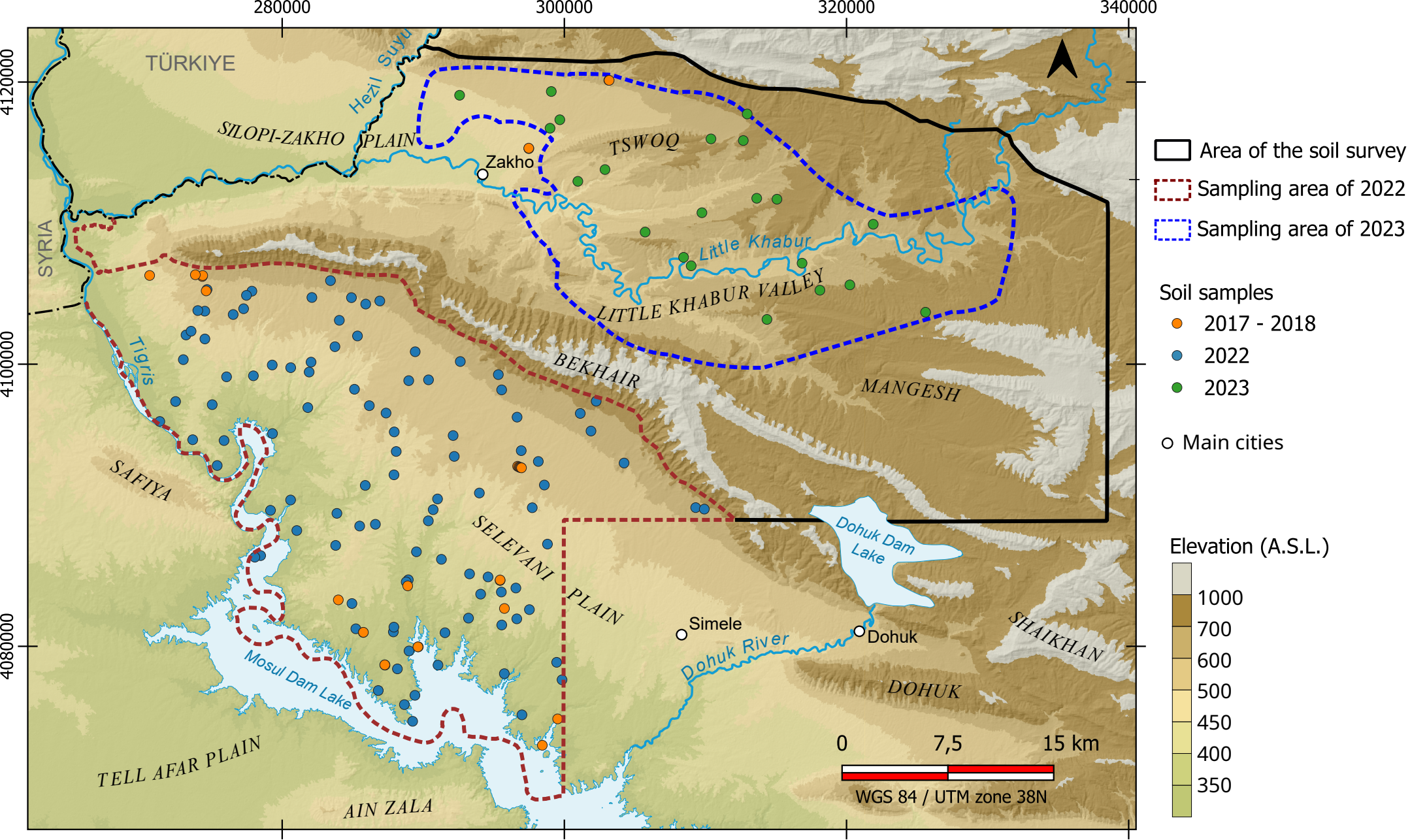
As for the table 2 we attached the SD row directly to the mean of both measured and predicted values with the "±" insert as you suggested.

The new section 3.1 is as follows:

" Observations from the different laboratory samples show a large variability in soil property distributions (Table 3, Figure 7). The pH values ranged from 6.93 to 8.2, with a mean of 7.3 ± 0.2, indicating a slightly alkaline soil environment. The CaCO3 content varied widely, from 3.61% to 84.27%, with an average of 30 ± 12%, suggesting significant differences in carbonate content across

the samples. Overall, only two samples had CaCO3 content below 10%, which indicates the strong relationship between soils and their parent material, mainly limestone, in this aridic environment. Total nitrogen (Nt) ranged 330 from undetectable levels to 0.67%, with a mean of 0.12 ± 0.1%, with the higher values concentrated in the upper 0 – 10 cm soil depth (20 on 20 of the highest measurements). While organic carbon (OC) content was generally low, with a mean of 1.1 ± 0.9%, the total carbon (Ct) content was higher, with a mean of 4.8 ± 1.7%, showing values approximately 360% higher. This pronounced difference between organic and inorganic carbon is well known for aridic and semi-aridic environments (Zamanian et al., 2016). Electrical conductivity (EC) values included some outliers above 500 µS/cm (n=5) - likly due to laboratory manipulation - explaining the high variability characterized by a standard deviation of 158 µS/cm. The mean weight diameter (MWD) of soil aggregates was higher mostly in the upper part of the soil profile (17 of the 20 highest values). Soil texture was predominantly silty-clay and silty-clay-loam, with average sand, silt, and clay contents of 19.4 ± 15.3%, 44 ± 10.5%, and 36.5 ± 14.4%, respectively. The upper depth increments (0 – 10 and 10 – 30 cm) showed slightly sandier textures than the lower layers (14 of the 20 highest measurements).”

**Citation**: https://doi.org/10.5194/essd-2025-418-RC5


280000
300000
320000
340000
4120000
4100000
4080000
TÜRKIYE
SYRIA
Hezil Suyu
SILOPI-ZAKHO PLAIN
Zakho
TSWOQ
Little Khabur
LITTLE KHABUR VALLEY
BEKHAIR
MANGESH
Tigris
SAFIYA
SELEVANI PLAIN
Dohuk Dam Lake
Simele
Dohuk
Dohuk River
SHAIKHAN
DOHUK
Mosul Dam Lake
TELL AFAR PLAIN
AIN ZALA
0
7,5
15 km
WGS 84 / UTM zone 38N
Area of the soil survey
Sampling area of 2022
Sampling area of 2023
Soil samples
2017 - 2018
2022
2023
Main cities
Elevation (A.S.L.)
1000
700
600
500
450
400
350