Response to Reviewer 1 Comments (RC1) for ESSD-2024-26

We thank you for your comments on our manuscript and suggestions for improving our work. We have addressed all the comments. Our response (AC) to each reviewer comment (RC) are shown in bold text below.

Best regards,

Anatol Helfenstein, on behalf of all authors

General comment:

The manuscript presents a high-resolution soil mapping platform developed for the Netherlands. The authors provide descriptions of how to perform the mapping, assess the uncertainty of the maps, and discuss the strengths and limitations of the mapping platform. They describe the software and computational network used for the mapping and share:

- most of the input data used for the mapping,
- all scripts used for soil and covariate data preparation, model training and validation, and
- the derived soil maps.

A comprehensive and detailed description is provided, which can serve as a guideline and be adapted for mapping soil properties elsewhere in the world. Therefore, the presented study is likely to attract significant international interest.

The manuscript's analysis is mostly clear, except for one aspect that pertains to the qualitative accuracy assessment. It would be important to clarify whether the BOFEK and BIS-4D maps used the same "National soil map of the Netherlands" as an input layer. If they did, then using BOFEK to assess the qualitative accuracy (patterns) of BIS-4D may not be plausible. However, BOFEK could still be used to better understand patterns visible on the soil maps, as done by the authors, providing reasoning for areas with specific properties such as higher sand content or lower bulk density, etc.

Additionally, it would be informative if the authors briefly explained why they did not add residual kriging after applying the QRF prediction.

Please find a more detailed review under "Specific Comments".

AC: Thank you for these general comments. We would like to clarify that while BOFEK used the entire national soil map of the Netherlands as a basis and starting point (https://doi.org/10.1016/j.geoderma.2022.116123), BIS-4D only used information about peat classes from the national soil map (https://doi.org/10.1038/s43247-024-01293-y, Fig. 5). Hence, we did not include any information about mineral soils from the national soil map as covariates (input layers) and therefore we maintain that comparing clay, silt and sand predictions with the BOFEK map is plausible and has an added value in addition to the quantitative, statistical accuracy assessment. However, the reviewer is correct in that there is some overlap for the peat areas, since in these areas, both the BOFEK and BIS-4D use information from the national soil map. We will clarify this better in a revised version of the manuscript by explaining that qualitative assessment of BIS-4D clay, silt and sand maps with the BOFEK map should focus on areas with mineral soils.

We chose not to use residual kriging for the QRF predictions because there was no spatial autocorrelation in the residuals (pure nugget effect in semivariogram). Furthermore, this would have significantly increased the computation time, especially for the soil properties with many observations (SOM and clay), and added an additional step to the methods, which are already quite complex. Spatial position was included in the model by including coordinates (Easting and Northing; Table 5) as covariates. We will add a sentence as follows at the end of section 2.5 (L247) to inform the reader: "We chose not to use kriging of the QRF prediction residuals (regression kriging) because there was no spatial autocorrelation in the residuals and to simplify the procedure."

Specific comments:

L80-93: please add reference about mapping activity in other European countries and give a short overview about those countries which were the first ones to prepare national coverage soil map.

AC: We appreciate the suggestion, but this would conflict with specific comment 3 of Reviewer 2 (David Rossiter), who even suggests that "The specific case of demand for soil data in NL is relevant, the paper could start there". There are many review papers of the history of soil maps in Europe (10.1016/j.geoderma.2013.05.003), digital soil mapping and the GlobalSoilMap initiative (e.g. 10.1016/j.geoderma.2021.115567), but this is beyond the scope here. We reference these papers in the manuscript in case the readers want more background knowledge.

L96: please provide exceptions: papers that present national 3D maps including several soil chemical and physical properties.

AC: These are provided in the review papers that we cite in this sentence from L95-96 (Chen et al., 2022 and Wadoux et al., 2021).

Figure 1: please add meaning of CLORPT in the figure's caption.

AC: We will add that CLORPT stands for the soil-forming factors climate, organism, relief, parent material and time in the figure caption as well as the reference (Jenny, 1941).

L127: … measured or estimated in the field at point locations … please consider to rephrase, because in the case of clay content, BD and SOM field estimation, too are used for building the model.

AC: We will adjust it to "measured or estimated in the field" as suggested.

Table 1: … Measured dry bulk density … is it correct?

AC: We prefer to not add the word "measured" as BD was both measured in the laboratory and estimated in the field.

L133: please recheck the appropriate meaning of O horizon, it might be a layer with undecomposed or partially decomposed organic material based on FAO's terminology or add reference of the horizon definition.

AC: We will adjust the sentence as follows to clarify: "We only included observations between 0 and 2 m depth excluding the O horizon, or the layer with dead plant

material, leaves, branches and other decomposing organic material on top of mineral soils."

L153: please add that year of sampling is given in Table 2.

AC: We will add this.

L157: Figure 2 shows locations of PFB, not BPK, please check it and correct text or figure caption accordingly.

AC: This refers to Fig. 2 of Helfenstein et al., 2024c, not Fig. 2 in this manuscript. We will change the text to "(Fig. 2 in Helfenstein et al., 2024c)".

L159: if you think that the skewness of some soil properties affected the model prediction, why didn't you transform those variables?

AC: Please note that this sentence is not only about skewness, as we describe in the sentence before that pH, sand and silt exhibit bimodal distributions. We state that distributions of the observational data affected model predictions because this is later discussed in the discussion section. We did not transform these variables because performance did not improve and to keep the model simple.

We argue in the discussion that the skewness of the data influences how mean and median predictions should be interpreted, not that the skewness of data leads to a decreased model performance (what the reviewer seems to suggest). E.g. positively skewed SOM leads to general overestimation of SOM on mineral soils when using mean predictions, median predictions are likely more valuable there (L333-338). In theory, you could indeed influence this by transforming the target variable to a more normal distribution. But when using RMSE as a metric to optimize this would likely lead to an increase in model performance (for mean predictions) on mineral soils and a decrease in model performance on organic soils. Overall, this would not lead to an increased model performance. In this context, it is not only about transforming variables but also about the choice in the metric to optimize (e.g. RMSE vs MAE).

L170: meaning of the sentence is not completely clear, does it mean that all samples were used for model training?

AC: We will rewrite the sentences in L170-172 as follows to improve clarity:

"For clay, silt, sand and CEC, no separate dataset with laboratory measurements was available for statistical validation, meaning all observations were used for model calibration. Therefore, statistical validation of these four soil properties was conducted using cross-validation of PFB laboratory measurements (Sect 2.6)."

L171: it would be clearer if the term "validation" were used exclusively for cases where independent data was available. Since independent data was not available in this instance, it would be better to use a different term than "validation".

AC: We disagree since adding an additional term instead of validation would be both confusing and unnecessary. 10-fold cross-validation is typically considered a validation strategy in statistical modelling and in digital soil mapping. During cross-validation, model performance is evaluated in each fold on data which was not used for model calibration.

Table 2: add the abbreviations used under column Method, i.e.: "Lab", "Field", where you explain the meaning of those.

AC: We will make sure to include the abbreviations in the caption as follows: "Lab = laboratory measurements; Field = field estimates; …"

L173-174: it would be informative to show sampling locations of LSK and CCNL dataset in the supplementary material.

AC: We will refer to Fig. 1 in Helfenstein et al., 2022 (10.1016/j.geoderma.2021.115659) for the LSK locations. As written in the text, the vast majority of LSK locations were revisited during the CCNL campaign, so these locations are almost identical and there is little added value in showing both. The supplementary material is organized by target soil property and already very extensive and the main paper also already has many tables and figures in the methods section. Therefore, we decided not to include an additional figure with the CCNL locations.

Figure 3: legend box could be reformatted, e.g.: centred, in one row and two columns.

AC: Thanks for the suggestion, we will do so.

L181-183: relationship between CCNL and LSK datasets is not clear please explain it a bit more.

AC: We will make it more clear as follows:

"The CCNL dataset consists of laboratory measurements from re-visited LSK locations in 2018, excluding locations that were no longer accessible. In contrast to LSK, during which soil samples were taken by soil horizon, CCNL locations were re-sampled at two fixed depth layers (0-30 cm and 30-100 cm)."

Table 4: … Obs. = number of observations …

AC: Thanks, we will adjust as you suggest.

L203: please add how you defined the 25 m resolution. Does it come from the density of soil profiles available for the mapping, or resolution of a covariate which is the most important for mapping soil properties? Or is there another reason?

AC: We will add that 25m resolution was chosen due to the resolution of the national land use maps (LGN).

L211-213: Recursive feature elimination (RFE) is highly computationally intensive in the case of >100 possible predictors. How did you perform it? After the de-correlation method how many predictors stayed for the RFE? Based on https://git.wageningenur.nl/helfe001/bis-4d/-/blob/master/31_regression_matrix_RFE.R?ref_type=heads script RFE was performed for SOM, clay and pH. Maybe it is not the latest script. If needed, pease clarify in the text for which soil property was RFE performed, how you handled the ones for which no RFE was performed.

AC: Indeed some further clarification could help make it more reproducible. We selected default values of covariates to retain (50, 30, 20, 15, 10), see L295 in the script you refer to. After de-correlation using the Pearson correlation coefficient, depending on the soil property, approx. 200 covariates remained. These were then subsequently reduced to 50, 30, etc. and the model with the best accuracy (RMSE) was used. This was done for all 9 target soil properties, not only SOM, clay and pH.

For those three, I merely noted the run time at the top of the script for benchmarking purposes. We will add the default values of covariates that we retained in the revised manuscript and mention that it is similar to the approach used in Poggio et al., 2021 (10.5194/soil-7-217-2021).

L215-216: please shortly describe why location-grouped 10-fold cross-validation was used, what the advantage of this method is.

AC: We will add the phrase that "Location-grouped cross-validation was chosen because observations from the same profile in both model training and validation can lead to overly optimistic model accuracy metrics".

L219: according to Table 2 only clay content is indicated as having both field estimate and laboratory measurement. If silt and sand content did not have field estimated values, please delete them here.

AC: The caption of Table 2 is "…soil point data used for model calibration". We tested whether including field estimates would improve model performance for all soil properties where field estimates were available (clay, silt, sand, BD and SOM). However, since performance did not improve when using silt and sand field estimates, they were not included in the final models and thus also not included in Table 2. We will keep the text as is since it is important to also inform the readership about what did not lead to model improvement.

L224-226: here again, the sentence starting with "For silt and sand …" is not in line with Table 2 and 3. Or maybe I miss something, please make it clear in the text and tables.

AC: Please see our comment directly above.

L236-238: please add more details to the sentence starting with "As we assigned …". The sample fraction is 0.8 in Table 6, isn't it the value used to divide out of bag fraction? Please explain why there were not enough samples for the out-of-bag.

AC: We will replace this last sentence (L236-238) with:

"When case weights are high, out-of-bag estimation is not possible because the observations with high weights are selected in the bootstrap sample of all trees, regardless of the sample fraction. Hence, we could not compute the out-of-bag error and use the permutation variable importance measure for these observations because they were never out-of-bag.

Table 5: please add resolution of the covariates.

AC: All covariates were resampled to 25m resolution (L203).

L251: in the case of PICP why 0.02 is the bottom threshold value?

AC: Thank you for catching this. Actually all quantiles between 0 and 1 were computed at steps of 0.02 (51st prediction interval – 49th prediction interval), not 0.01. Therefore, it makes sense that 0.02 was the bottom threshold value. We will correct this in the text (L250) as follows: "…, all quantiles from 0 to 1 at steps of 0.02 were predicted…".

L258: why do you call external accuracy assessment the one that you compute on PFB dataset? PFB was used to calibrate the model, therefore would belong to internal accuracy assessment.

AC: We disagree with the reviewer on this matter. Cross-validation is a model-independent and thus external accuracy assessment method. It can be used regardless of the model used for calibration and prediction. Please see also Table 3 in Helfenstein et al., 2022 (10.1016/j.geoderma.2021.115659).

L317-318: please discuss the reason for having low accuracy for P-oxalate.

AC: We explain several reasons for low accuracy of P-oxalate in paragraph 2 of Sect. 3.3.2, which makes more sense for the overall structure of the paper. This first part of the results should not go into too much detail for one soil property but rather provide an overview of the results.

L319-321: can it be the reason that there is some difference in data quality or measurement accuracy or the way that a measurement method is performed (even if it is done with the same methodology there can be some difference in how the sample is pre-treated, etc.) that in PFIB dataset the accuracy of BD and P-oxalate is higher than in the LSK dataset?

AC: Although this is a good point, there is no evidence that suggests that pre-treatment or measurement method was different. Instead, there is a lot of evidence to suggest that it rather has to do with the different sampling designs of the two datasets (L319-321).

L322-325: it would be informative to compute relative error which would support further explanation.

AC: We agree, but we did not want to introduce an additional metric simply for this single explanation (1 sentence). We think that the explanation is clear and understandable also without computing relative errors.

Table 8: What can be the reason for higher MEC for 15-30 cm depth in the case of BD and pH? Please discuss possible reason of decreasing accuracy with depth in the case of sand, silt, and clay content around L350.

AC: We agree that it is good to add a possible explanation on this. We will add the following in L350, after it is explained that accuracy decreased with increasing depth:

"Deeper soil layers are generally more difficult to predict because limited information about the subsoil can be derived from most covariates, especially remote sensing products. However, for BD and pH, the accuracy from 15-30cm depth may have been higher than from 0-15cm depth because only 245 observations were available for statistical validation in LSK from 15-30cm depth (Tables S4 & S6). Therefore, the metrics computed via design-based inference from 15-30cm depth for BD and pH are likely less representative of map quality compared to metrics of the other depth layers, where many more observations were available."

L345-347: please add possible reasons for having higher uncertainty of the maps in river and Pleistocene areas.

AC: As the topic sentence of this paragraph explains (L339-340), uncertainty was high when mean and median predictions fell within a range with limited calibration data. "…given its bimodal distribution, the uncertainty for sand was highest in areas where predictions ranged 345 between 25 - 75 % (for example in the river areas) and uncertainty was comparatively low in marine clay areas (< 25 % sand) and Pleistocene areas (> 75 % sand) (Fig. 4c, g & k)" as written in L345-347. Note that thanks to the comment, I have realized there was a mistake in this sentence.

Uncertainty was low in both marine clay areas and Pleistocene areas (and not only in marine areas), as the map in Fig. 4k clearly shows and as supported by Fig. S30 in the supplements. In the above quotation, we have corrected this. In summary, there was higher uncertainty in river areas because sand content was often between 25-75% (in this range there was little calibration data), while uncertainty was low for marine and Pleistocene areas (in those ranges there was more calibration data).

L366: "avoid presenting overoptimistic results to end users", maybe it could be added that it is important to clarify the quality of data used for the mapping for the users of the derived maps. Just a note: uncertainty of estimated input data is higher than uncertainty of input data measured in the laboratory.

AC: We agree with the reviewer that it is important to clarify users about the quality of the input data, but here we are arguing that accurate quantification of the prediction uncertainty is essential (L365). Therefore, we would like to keep the text as it is.

L371-372: is there any overlap between the predictors of BIS-4D maps and information considered to prepare the soil physical units map? If both considered peat classes from the "National soil map" or and groundwater classes or land use map of HGN, LGN or geological units or geomorphology, etc. "patterns" of the maps will be similar. Please consider it and rephrase the paragraph if needed. If BOFEK is the same national soil map as indicated in Table 5 for the peat classes, then comparison of BOFEK and BIS-4D is not plausible.

AC: Please see our first comment (AC) above. We will add the following text in the methods to explain the overlap but maintain that it is plausible to compare BIS-4D maps with BOFEK:

"Note that we did not compare visual patterns of the national soil map (de Vries et al., 2003) and the soil physical units map (BOFEK; Heinen et al., 2022) to BIS-4D predictions in peat areas, as covariates of peat classes were used in model calibration (Table 5 and Fig. 5 in Helfenstein et al., 2024c)."

L378: please add proportion of peatlands areas to clarify the size of the area affected.

AC: Good point. We will do so: "…should be used with caution in peatlands (approx. 15% of the surface area), …"

L389: It is not clear why mosaicing created artificial lines, please describe it more or state that more analysis is needed in the future to solve it.

AC: We have adjusted the text as follows to improve the explanation:

"Other artifacts were due to the combination of several Sentinel 2 images from different days in one month to obtain one monthly, cloud-free mosaic (Sect 2.2). Image mosaicing created artificial lines due to alterations in the brightness, hue and colors from images of different days."

L503: … frequency of agricultural machinery on the fields …

AC: We will adjust the sentence as you suggest: "For example, BD is strongly dependent on the size and driving frequency of agricultural machinery on the fields (Stettler et al., 2014).

L510: please consider reasons coming from management – e.g.: typical depth of fertilization – and add it in the text as further possible explanation.

AC: We will add the following sentence at the end of L510: "P from fertilization largely stays in the upper soil layers."

L512: There are some papers on mapping soil phosphorus content at high resolution, please compare your results with those.

AC: We will add supporting literature: "…likely due to missing (historic) management data. This is supported by other studies of mapping soil P content at high resolution (Delmas et al., 2015; Matos-Moreira et al., 2017; Kull et al., 2023).

L550: … clay, silt, and sand content, SOM …

AC: SOM is also a content (mass percentage). Therefore, we will correct the sentence as follows: "For example, maps of clay, silt, sand and SOM content can provide…".

L552: … more interested in more complex …

AC: Thanks for catching this mistake. We will correct it as follows: "are mostly interested in more complex soil information,…".

L564, 577: … clay and sand content and pH …

AC: Thanks for catching this mistake. We will correct it as follows: "These recommendations hold true especially for clay and sand content and pH, which…".

L578: the BD maps of the BIS-4D were not among the most accurate ones based on the design-based inference, please rephrase the sentence.

AC: Good point. We will adjust the sentence as follows:

"This is mostly in agreement with Chen et al. (2022), who found that pH was the best predicted standard GSM soil property, and PSFs (i.e. clay, silt and sand) were predicted third best, based on a review of 244 articles."

Supplementary material:

- please add meaning of variables shown on variable importance plots, maybe as a table somewhere before Figure S11,
- please format page S37.

AC: This information is included in table 5, where the names of the covariates are provided. Furthermore, the openly available code contains readme files for every single covariate used in BIS-4D, which allows users to easily get the covariate metadata (e.g. https://git.wur.nl/helfe001/bis-4d/-/blob/master/data/covariates/geology/geomorph2008_genese_25m_readme.txt?ref_type=heads). Lastly, this information is also part of the covariates dataset provided as an asset alongside the paper (https://doi.org/10.4121/6af610ed-9006-4ac5-b399-4795c2ac01ec).

However, to make this clear, we will add the phrase: "Covariate names from the y-axis can be found in code (https://git.wur.nl/helfe001/bis-4d/-

/tree/master/data/covariates?ref_type=heads) and covariate dataset (Helfenstein et al., 2024b).”

We will format page S37 in the supplements as suggested.