We thank anonymous Referee #2 for valuable suggestions and comments, which have greatly contributed to the enhancement of our manuscript. Our responses are provided in black text color following each comment and suggestion by the referee in blue text color:

Overall evaluation:

- I feel that the paper is a great effort by the authors to draw together a set of soils data for Ethiopia and improve the spatial resolution of the mapping. I think just pulling together the data set is a big achievement.

Response 1: Thank you for the positive feedback and compliments on our work

- However, I feel the paper lacks a critical evaluation of the results and of the subsequent learning and recommendations that could be made. To do this it needs an assessment of where the modelling worked well and where it didn't and explanations of why these results may have occurred.

Response 2: Thank you for the comment. The modeling accuracy was assessed based on the standard cross-validation technique which involves the overall map accuracy. It is a resource and time-demanding (which also was not the scope of the present study) to consider model-free and design-unbiased accuracy assessment which is believed to be achieved with probability sampling, while taxonomic correctness is one of the key determinant factors to be considered in such class/Reference Soil Groups (RSGs) mapping.

Digital soil mapping (DSM) product users have indicated critical concerns to what degree DSM products represent the actual soil landscape spatial patterns, as similar/close quantitative accuracy statistics might show different soil class spatial patterns. To address this concern, we employed an expert-based qualitative assessment of the model output. This technique was used to complement model-based accuracy assessment and confirm/indicate where the modeling specifically worked well and where it didn't. This was implemented by a panel of senior soil specialists/pedologists checking the map based on objectively selected geographic windows across Ethiopia, representing different agroecological zones known to have diverse soil occurrences, and familiar to the panel of experts. Accordingly, the outcome of the evaluation which is an indicator of the model performance across geographic windows presented interms of aggregated ratings (lines 229 and 230): 1. confirmed with '*no concern'*, 2. confirmed with

"*minor concern"*, and 3. confirmed with '*major concern'*. However, we accept the comments and we will elaborate on the findings of the qualitative evaluation as per pedological-based interpretations/assessments both in the examined geographic windows and prominent contrasting landscapes of Ethiopia.

To provide some reflection on the basis of spatial windows, for instance, in the northeastern lowlands of Ethiopia, mainly along the "Denakil" depression, it is observed that the model overestimated Fluvisols; and confused Fluvisols with Vertisols. Further, mainly Solonchaks, believed to be peculiar features of that particular landscape and Leptosols are under-represented. In some parts of the southeastern lowlands of Ethiopia, Calcisols spatial distribution is under-represented and Cambisols were overestimated. The modeling didn't work well in these cases which may be attributed to the low number of soil profile observations (Figure 5) in those areas. This implies that we need additional soil profile observations. The above discussion will be added in the revised version under the new heading **3.4. Evaluation of results and future direction.**

- I think the discussion of the maps with experts is a really useful way of validating the maps and more could be made of the results of these discussions.

Response 3: We accepted the comments, we will add more soil-landscape-based elaborations (kindly see Response 2) based on examined geographic windows and well-known national spatial patterns, as the team involves a panel of senior soil surveyors/experts/pedologists who have been involved in many soil survey and mapping missions across a mosaic of Ethiopia's landscapes.

- There needs to be a discussion about where results are unexpected/expected and how that links back to figure 5 and the availability of the input soil profile data and covariates in different areas.

Response 4: Thank you for this comment, we will address it (kindly see also Response 2). There are areas where fewer soil observations (explained in lines 285 to 287) and sparse geographical coverage affect the modeling performance. This was observed and reported by the panel of experts zoomed-in assessment across areas labelled as 'minor' and 'major' concerns and across some landscapes such as in the eastern lowlands. Besides, geographic coverage of quality input

soil profile data, adequate representation of the feature space could affect the model performance. Sometimes given the covariate issue and examining spatial details relatively similar, some unexpected spatial patterns might be due to issues related to the adequacy of representing the feature space. In addition, the granularity, level of detail and quality of the covariates towards the model performance will be further elaborated, in such a way as to highlight areas that are worth consideration for future similar studies and efforts to improve the map accuracy.

- The paper needs to highlight what we can learn from mapping in Ethiopia for mapping in similar landscapes. If this can be added I think it would be a really valuable addition to e DSM literature.

Response 5: One of the key insights gained from this study is the critical role of collating existing soil profile data. It is important to recognize that conducting repetitive soil characterization and classification exercises or an effort to update existing legacy soil maps through new soil survey campaigns can be both costly and time inefficient. Similarly, for countries like Ethiopia which are very vast and characterized by diverse soil forming factors and soil resources, a conventional mapping approach would be much more resource and time-demanding. Therefore, it is imperative to explore alternative approaches that maximize the utilization of available and optimal soil profile data and digital soil mapping techniques which the paper aims to address.

In addition, addressing the issue of data standardization within data collation methodologies is of utmost importance. By establishing standardized data collection practices, we can ensure the compatibility and comparability of collated data for effective utilization in digital soil mapping (DSM) models throughout Africa. The paper emphasizes the significance of implementing data collection standards and practices in Ethiopia and other Sub-Saharan African regions. This will enable the generation of a sufficiently large number of observations, which are essential for developing data-driven DSM models and other precision agronomy applications.

It is essential to note that the recommendations presented in this paper extend beyond Ethiopia's borders and hold relevance for other countries in Sub-Saharan Africa. These recommendations provide valuable insights and guidance for the adoption of standardized data collection practices across the region. By embracing these recommendations, researchers and practitioners can ensure the generation of high-quality data, thereby facilitating the development of robust and effective DSM models and precision agronomy approaches. Some of these learnings will be added and discussed in the revised manuscript.

Specific queries:

- Could the resolution of the input data explain why the results may not be as expected in certain areas?

Response 6: Yes, among other factors, if we have separately examined the effects of the covariates, the spatial resolution and level of detail could contribute to why the results are unexpected in certain areas. For instance, within the given spatial level of examination, the sequence of some RSGs showed different patterns which could be captured by better resolution parent material map in the SCORPAN model. We will highlight this issue in the revised manuscript.

- In the discussion of the confusion matrix (Table 1) the authors could look at where there are large differences between soils pedologically and where a miss mapping of soils might lead to different management decisions in areas.

Response 7: Thank you for raising this issue and for the comments. In the confusion matrix (Table 1), the quantitative classification errors (omission and commission errors) need to be interpreted/checked in terms of the soil's pedological similarity/differences which is commonly called 'taxonomy distance'. It is such an evaluation that will add value to interpreting the errors from producers' and users' perspectives and check areas of concern to implement management decisions. In soil class mapping where classification accuracy is represented by a confusion matrix, literature indicated, it is likely that not all errors are equally serious. Some errors are more serious than others in terms of soil properties, soil-forming process, ease of map making and application of the map. For instance, from the user's perspective, Vertisols predictions were distributed to incorrect Leptosols and Nitisols classes which implies leading to significantly different management decisions in terms of soil depth, aeration, and acidity. The same applies to miss mapping of Arenosols as Luvisols and Vertisols. The miss-mapping interpretation needs to be supported based on the soil's taxonomic distance, which determines class similarity and dissimilarity determining different management decisions and hence, implying, fractional recognition needs to be given to some incorrect allocations represented in the confusion matrix.

- The paper mentions a rerun of the modelling after the workshop. Can the authors explain what was changed to improve the results between the 2 runs and which versions of the runs are presented in this paper.

Response 8: After re-running the model, about ten soil scientists and geospatial experts (lines 242 and 243) re-evaluated the output using districts selected based on the feedback from the first review, which was mainly on areas where there was "minor" and "major" concerns. For instance, in areas where Vertisols, Fluvisols, and Leptosols were reported to be overestimated, improvements were observed. Further, underestimated RSGs (Alisols, Solonetzs, Planosols, Acrisols, Lixisols, Phaeozems, and Gleysols) showed slight area coverage and pattern improvements. However, the total area for Leptosols and Cambisols increased from the first run due to the partial exclusion of the mask layer used in the first round modeling effort. The mask layer used in the first run was criticized for quality issues as it excluded significant soil areas and its limitation to capturing non-soil areas such as rock outcrops/rocky surfaces, salt flats, swamps and sand dunes across the different landscapes. Nevertheless, the spatial patterns of these soils occurring across previously considered "non-soil areas'' were examined by the panel of experts. In parallel, geospatial and soil experts checked the raster map of the RSGs in the GIS environment to ensure areas with 'no concern' before re-running the model are kept the same or changes are accepted by the panel of experts. The map from the second run is presented in this paper.

- I think its structure needs some thought specifically. The results of the validation described in section 2.4.2 need to be part of the results rather than the methods.

Response 9: Thank you for the comment. In section 2.4.2. we presented how we did the qualitative validation procedures (i.e. expert evaluation) and the outcome of this process is presented in the result section (sec 3.3). We thought this flow was much easier to follow the paper. Therefore, we kindly ask the reviewer to allow us to maintain the current structure of these sections.

Points of clarification:

- Line 59: What is meant by "hardly available"

Response 10: As elaborated for Referee 1 (See Response 4 of AC 7) we wanted to say that a national quantitative and spatially continuous predicted reference soil group/soil type map does not exist. We admit that hardly available is confusing and in the revised manuscript, it is revised by "does not exist".

- Line 113: What criteria were used to define if a profile is complete and clean?

Response 11: The criteria used were basic profile information/data required for classification of RSGs. For clarity, the statement will be amended as: ..... cleanness, i.e., profile points with basic data/information for classification of RSGs.

- Line 223: How were the polygons for review selected?

Response 12: In order to represent every part of the country, the polygons/geographic windows for qualitative assessment were purposely selected by a panel of senior soil specialists/pedologists/soil surveyors before breakout sessions and proceeded to group works. The revised version will be updated by adding the phrase "purposely". The experts were drawn from different corners of the country and had been involved in different soil survey missions across Ethiopia. Hence, each suggested geographic window was debated and agreed upon based on soil diversity, contrasting/unique soil-landscape relations, availability of familiar experts in the panel, and agro-ecological zone coverage.

- Line 233: How are the authors looking to improve the version of the map from the first version?

Response 13: Thank you for raising this issue. The first version of the map will be improved by ensuring additional input profile data from under-represented geographic and feature spaces, and covariates with improved resolution, quality and level of detail including through the implementation of different covariate selection procedures. Application of a robust modeling technique that accommodates neighbourhood size and connectivity analyses requires due consideration by future studies. It is also recommended to implement unbalanced data treatment and de-clustering techniques to overcome issues likely to arise from class imbalances and biased datasets in such kinds of soil class/type mapping efforts. The above statement will be added in the revised version under the new section, 3.4. Evaluation of results and future direction.

- Line 247 – 253: Do the number of samples used represent what would be expected in terms of areas of specific soils in Ethiopia or are the input data biased to specific land cover or soil types.

Response 14: In general, ignoring the temporal resolution, i.e., from the 1970s to the 2020s, the number of samples is expected to cover areas of important agroecological zones and land use/covers. However, in terms of areas of specific soils of Ethiopia, while the 1st, and the 2nd largest input data were from Vertisols and Luvisols, their relative area coverages were in 3rd and

6th positions, respectively. This bias might have happened because of the soil survey interests. For example, many surveys focused on Vertisols and Luvisols for the purpose of agricultural intensification/mechanization and irrigation in areas where these soils are situated. This signifies the need to focus on future soil data collection to consider soils with fewer input data compared to their relative area coverage. Moreover, this study utilizes the most extensive soil profile observation data available to date for the generation of a comprehensive soil-type map of Ethiopia. Despite the inherent uncertainties associated with data representation, this is the first significant endeavor based on such a large-scale observation effort. This description will be added to the revised version under the new section 3.4. Evaluation of results and future direction.

- Line 274-278: Do the authors see a difference in the quality of the results where they had an increased density of input profiles?

Response 15: In general yes, but not in all the cases, for instance, based on geographic and feature space coverage and RSGs diversity.

- Figure 6: Add an axis label to the X axis

Response 16: Thank you for the comment. We will label it.

- Line 409-418: The authors need to discuss in more detail the reasons why certain points in the topographic sequences do match other work and where they don't and offer potential explanations of why.

Response 17: Thank you; we will elaborate further as suggested.

- Line 428-435: This section assumes that the new soil grids that have been generated are better than the "soil grids" without explaining what the insight comes from the new modeling and why it's important. It would also be valuable if the authors could offer insight into which of the 3 reasons the results may be different.

Response 18: Thank you for the comment. We will elaborate further. Kindly please note that we based our comparison on the reported map accuracies, implementation of expert-based qualitative assessment of spatial patterns, and number and distribution of input soil profile observations. We will elaborate more and recommend the need for quantitative comparisons of legacy soil maps (including "soilgirds") in terms of how well they represent soil geography. Hence, users will get insights into the applicability of various DSM products at different spatial scales and geographic windows.

- Line 441-444: Is it likely that the data used in this study are biased and can the authors offer a recommendation on what new data might be needed in which areas to improve the results.

Response 19: Part of this query is addressed in the above (kindly see Reference 14). Keeping the temporal resolution constant, as the data source between the 1970s and 2020s, the input data are biased to specific land uses (cultivated/arable and grazing lands) and agroecological zones of Ethiopia (see lines 290 to 301). Hence, additional legacy data are required from less represented land uses such as forests, shrubs and bushlands. However, in some geographic areas such as the north and southeastern lowlands and in some agroecological zones where there is no/under-representation of input data, additional new data are required from more land uses.

- Lines 473-479 it is unclear whether the rerun version of the map is what has been presented in the current paper whether that is something that is to follow. If it isn't presented can the authors explain why not.

Response 20: Thank you for the comment, we will elaborate further. This query is addressed in the above (kindly see Response 8). The map from the second run is presented in this paper.