Reviewer 5:

The production of global maps of SOC stocks is valuable research and the dataset is of high interest for the global community. This manuscript is therefore timely. The current methodology and datasets used to build the maps, however, have several very fundamental issues that need to be resolved before any publication is made. For several of these issues, well-accepted solutions exist in the literature, and it is not clear if the authors are proposing something new (in which case, it should first be tested), or if they simply decided not to account for past developments. The manuscript suggests an important lack of familiarity with the current state of the art in soil science and digital soil mapping. All of the above is common knowledge and not prone to any specific discussion in the field.

I focused on the very major comments because I see that the other reviewers mentioned several minor issues.

The key issues are (explained in more details below):

1. Predictions should be made for depth intervals, not exact depths.
2. Depth harmonization (e.g. mass-preserving splines) is required to address differing sample supports.
3. Input datasets (e.g. RaCA, Australia) require quality checks and better sources.
4. Clarification needed on coarse fragment data and SOC measurement methods.
5. Modeled data should not be mixed with measured data.
6. Pseudo-observations should be added for areas with no SOC (e.g. deserts).
7. The choice of 100 m resolution should be justified.
8. Cross-validation, not single data splits, should be used.
9. Bias correction is unusual and needs justification.
10. Uncertainty assessment should use prediction intervals (e.g. quantile regression forest).
11. Check for residual autocorrelation and consider kriging residuals if needed.

Author Response:

Thank you for your review. We appreciate the time that you took to provide feedback on our work.

Depth intervals. In soil science, predictions are made for depth intervals, not specific depths as the authors did here for 30 and 100 cm. The concept of calculating a stock at exactly 30 cm or 100 cm depth does not make scientific sense. A stock is based on volume, i.e., a depth interval. The authors justify this in Section 2.1 by stating “… continuous soil depth functions

(Malone et al., 2009), which have shown significant variability in results." Using depth functions is the standard in soil science because soil samples are collected at different depth intervals (e.g., 0–10 cm or 5–15 cm) and need to be harmonized before use. This is a necessary pre-processing step. It becomes very clear in the next sentence of the same section why it is needed: "At a depth of 30 cm, we included 84,880 ground truth data points." What does this mean exactly? Did you merge together samples collected for any interval between 0 and 30 cm? What did you do for samples that were collected, for example, on the interval 0–40 cm? This shows that using depth harmonization is a prerequisite as much as calculating the stock for a depth interval. The justification given in Section 2.1.4 for depth harmonization does not make sense. I invite the authors to read the papers on depth harmonization in the soil science literature, where the approach given here seems outdated and not correct.

Further on this matter of the depth interval, the authors are currently mixing samples with different supports. It is not the same to estimate the SOC stocks for a sample obtained on a 0–5 cm interval compared to a sample with a 0–30 cm depth support. These are very different, yet the authors are putting everything together and considering it the same measurement. This again supports the need for using a mass-preserving spline as a basic pre-processing step for SOC stock data.

Author Response:

Thank you. To address these concerns, we have added further clarification to the methods **Section 2.1.4,** as follows:

"Many of the datasets were already standardized to the 0-30 cm or 0-100 cm depth intervals. For profiles with multiple layers, we only considered those with both upper and lower depth boundaries reported and combined consecutive nested layers to calculate total SOC. For the 0-30 cm interval, we included profiles that adequately covered the topsoil layer, allowing a small tolerance to maximize dataset inclusion (20-45 cm actual soil depth) while maintaining depth consistency. Profiles that were too shallow or had gaps in measurement were excluded. For deeper layers (0-100 cm), we similarly retained only profiles with complete depth coverage, again applying the small tolerance for uniformity (≥80 cm actual soil depth). Linear interpolation was applied where minor adjustments were necessary. Peatland datasets were handled separately to retain deeper measurements due to limited data availability. Samples missing depth information were excluded, and duplicate entries with identical coordinates and SOC values were removed.

To evaluate the quality of the harmonized datasets, we tested their predictive performance using biome-specific models, training each model only on samples from the corresponding biome. We calculated R^2 values for biome-dataset combinations with more than 50 samples. Globally distributed datasets such as ISCN, WoSIS, and ISRIC-WISE were

the most widely represented and consistently showed high predictive reliability. At 0-30 cm, WoSIS performed best in tropical biomes (average test R^2 = 0.62), while ISCN showed broad coverage with good performance in temperate biomes (average test R^2 = 0.45). Similar trends were observed for the 0-100 cm depth interval, though performance variability increased with depth and region, with some datasets providing stronger signals in specific biomes. This assessment indicates that the harmonized datasets reliably capture SOC patterns at the biome level, supporting their use for biome-specific modeling and prediction."

Problems with input point datasets. There are several known problems with the datasets the authors used. Most of the US data are based on the Rapid Carbon Assessment (RaCA) dataset. This dataset is known to have several issues. First, most values in this dataset are not measured, but predicted with infrared spectroscopy, which introduces an additional source of error. Second, there is a high likelihood of issues with bulk density measurements. The authors should contact the dataset maintainers to get more information. This dataset should at least be checked carefully and harmonized to retain the highest-quality bulk density measurements. As an example, the best mapping paper so far on soil properties in the US did not incorporate the RaCA dataset as input (https://acsess.onlinelibrary.wiley.com/doi/10.1002/saj2.20769).

Author Response:

Thank you. The RaCA dataset represented a relatively small portion of the training data for the biome models. At 30 cm depth, RaCA contributed roughly 15% of the data in temperate grasslands, savannas and shrublands, as well as in temperate coniferous forests, about 10% in temperate broadleaf and mixed forests, and in deserts and xeric shrublands, and minimally (1-4%) in four other biomes and absent in all remaining biomes. At 100 cm depth, it contributed ~20% in temperate coniferous forests and in deserts and xeric shrublands, ~15% in Mediterranean forests, woodlands and scrub, in temperate broadleaf and mixed forests, and in temperate grasslands, savannas and shrublands, and minimally in two other biomes.

We found that the RaCA dataset provided moderate yet positive predictive performance, reflecting meaningful spatial structure. Across the relevant biomes, the average performance (test R^2) for RaCA was approximately 0.25 at 30 cm depth and 0.31 at 100 cm depth. Based on this performance, we decided to include the RaCA dataset in the models.

We have added this clarification to **Methods section 2.1.1**.

Another concern is the lack of real data for Australia. The authors randomly sampled an old map to generate data (note that this 2014 map has been updated in 2022 with entirely new predictions). This is problematic. There is a wealth of publicly available datasets in Australia

that the authors could use, available through the SoilDataFederator. Alternatively, the authors could contact the authors of the recent paper on mapping SOC stocks in Australia to obtain their pre-processed data.

Author Response:

Thank you for your comment. SOC in Australia is derived from a combination of ground-truth points, biome-specific models, and stratified map-based sampling. We have provided further clarification in **Methods Section 2.1.3**., which now reads:

"Soil organic carbon in Australia was derived from a combination of ground-truth points, stratified map-based sampling, and eight globally applied biome-specific models. For Australia, our dataset included recent ground-truth SOC data from WISE30sec v.3 (Batjes, 2016), WoSIS (Dec 2023), ISCN v.3 (Nave et al., 2022), MarSOC (Maxwell et al., 2023), and WHRC-TNC (Sanderman et al., 2018). Using the quantile-based bin approach previously described, we additionally sampled 500 points evenly across the full SOC range of the Australian SOC map developed by Viscarra Rossel et al. (2014). These 500 samples accounted for no more than 3% of the eight biome datasets covering Australia, providing regional detail without dominating the broader analysis."

Two major attention points are raised with the SOC stock calculation:

Coarse fragments are notoriously difficult to obtain and predict. How did the authors obtain these data? It is not mentioned anywhere except for a brief statement that "it is not available everywhere." This needs to be made very clear because it will substantially affect the SOC stock calculation.

Author Response:

Thank you for this comment. To address it, we have added the following clarification to the **Methods Section 2.1.4** of the manuscript:

> "Across datasets, coarse fragments were inconsistently reported, but in most cases they were either explicitly provided or implicitly accounted for during SOC stock calculations. Many harmonized datasets that reported final SOC stocks did not include CF values as a separate variable, but indicated that coarse fragments were handled internally. Datasets that did not provide final SOC stocks often included CF measurements, although these were only partially available (i.e., not all data points). We adjusted SOC stocks to account for the presence of CF when possible. Overall, coarse fragments were generally considered in stock estimation, but differences in their treatment remain a source of uncertainty in global SOC datasets (Hengl et al., 2017; Poeplau et al., 2017), and ongoing efforts aim to improve consistency."

The authors ignored the difference between SOC obtained by dry combustion and the Walkley–Black method. They cite one paper to justify this, but the difference between methods can be significant. It is standard to apply a correction or otherwise account for this difference.

Author Response:

Thank you for this comment. To address it, we have added the following clarification to **Methods Section 2.1.4** of the manuscript:

> "Soil organic carbon was assessed using different methods (e.g., dry combustion or Walkley-Black) across datasets, reflecting the extended timeline over which data were collected. Most datasets compiled in this study had already been internally harmonized using conversion factors, and we did not apply further adjustments for potential method-based discrepancies. Most datasets reported final soil carbon stocks standardized in t C/ha. Older campaigns relied more on loss on ignition (LOI) and Walkley-Black methods, potentially introducing regional differences despite conversions. In contrast, more recent datasets, which we integrated to enhance spatial coverage, were predominantly analysed using dry combustion. Thus, methodological differences remain a potential source of bias in the compiled SOC estimates; however, remaining method-related effects are expected to be relatively minor compared with the overall spatial variability captured by the model."

Mixing measured and modeled data. The use of map sampling to generate observations for model fitting is a serious problem. The authors have combined observations from measured SOC stocks (derived using various calculation methods) with values obtained from SOC stock maps. This is problematic, as the map values are already modeled predictions, not direct measurements. They are smoothed and often carry significant uncertainties. The authors should avoid this practice.

Author Response: We thank the reviewer for highlighting this point. We have added detail regarding the inclusion of subsampled data in **Section 2.1.3** of the methodology, which now reads:

"We analysed model performance both with and without subsamples derived from the previously described gridded datasets. The results show that, when subsamples are excluded, model performance remains largely robust, with only minor reductions of 0.01-0.02 in R^2 values across most biomes and ecosystems (Table S8). Montane grasslands are an exception, showing a larger drop in R^2 due to the low number of samples in this biome. At 100 cm depth, removing subsamples similarly preserves model performance, with only marginal changes of 0.02-0.03 in R^2, again with the exception of montane grasslands. These

results indicate that, although map-derived samples contribute to model training, they do not fundamentally alter overall model performance or spatial patterns. The model is therefore not overly reliant on these potentially biased samples. We include these samples to improve spatial representativeness, but we have verified that their inclusion does not introduce substantial bias in our predictions. The full biome-level results with and without subsamples are available in Table S8."

	with subsamples		without subsamples	
Biome/Ecosystem soc30	**Full Count**	**Full R^2**	**Full Count**	**Full R^2**
Forests				
Tropical and subtropical moist broadleaf forests	9,830	0.84	9,162	0.82
Tropical and subtropical dry broadleaf forests	3,438	0.74	3,424	0.74
Tropical and subtropical coniferous forests	1,234	0.79	1,232	0.79
Temperate broadleaf and mixed forests	24,935	0.72	24,889	0.73
Temperate coniferous forest	8,820	0.71	8,781	0.71
Boreal forests / taiga and tundra	4,595	0.60	No subsamples	
Mangroves - based on Global Mangrove Watch (2020) extent	1,577	0.80	1,238	0.79
Grasslands and shrublands				
Tropical and subtropical grasslands, savannas and shrublands	4,074	0.83	3,439	0.83
Temperate grasslands, savannas and shrublands	8,667	0.74	8,448	0.74
Flooded grasslands and savannas	1,298	0.76	1,024	0.76
Montane grasslands and shrublands	418	0.73	33	0.40
Other biomes				
Mediterranean forests, woodlands and scrub	7,042	0.79	6,962	0.79
Deserts and xeric shrublands	9,041	0.70	8,493	0.70
Peatlands - based on UNEP classification (2022)	3,372	0.67	2,805	0.67

	with subsamples		without subsamples	
Biome/Ecosystem soc100	**Full Count**	**Full R^2**	**Full Count**	**Full R^2**
Forests				
Tropical and subtropical moist broadleaf forests	2,155	0.90	1,573	0.87
Tropical and subtropical dry broadleaf forests	1,322	0.71	1,295	0.71
Temperate broadleaf and mixed forests	14,516	0.72	14,503	0.72
Temperate coniferous forest	6,919	0.70	6,873	0.70
Boreal forests / taiga and tundra	1,716	0.68	No subsamples	
Mangroves - based on Global Mangrove Watch (2020) extent	1,453	0.80	1,097	0.78
Grasslands and shrublands				
Tropical and subtropical grasslands, savannas and shrublands	1,406	0.90	1,360	0.90
Temperate grasslands, savannas and shrublands	8,236	0.62	8,219	0.62
Flooded grasslands and savannas	1,034	0.77	790	0.75
Montane grasslands and shrublands	385	0.56	21	0.28
Other biomes				
Mediterranean forests, woodlands and scrub	1,298	0.83	1,297	0.83
Deserts and xeric shrublands	4,431	0.66	4,370	0.66
Peatlands - based on UNEP classification (2022)	1,936	0.77	1,286	0.75

Table S8. Model performance (R^2) and sample count per biome and ecosystem for soil organic carbon (SOC) predictions for 0-30 cm and 0-100 cm depths. Results are shown for biome-specific models trained with the full dataset and with or without the inclusion of subsampled gridded datasets.

Ignoring zero-SOC areas. The authors ignored areas that contain no SOC in soils, such as deserts. It is common procedure to add pseudo-observations in such areas to prevent the model from predicting SOC stocks where there should be none (because of smoothing effects) and to reduce uncertainty in total SOC stock estimates.

Author Response: We confirm that we did not include value-zero pseudo-observations in our dataset. Many desert regions are already represented in global and regional datasets with very low or value-zero SOC. We modeled deserts and xeric shrublands separately, which further limits the risk of over-prediction. The combination of these observed data and Landsat surface reflectance inputs allows these regions to be accurately represented in our global SOC map.

The specific resolution chosen (100 m) is not justified. We know that it is simply a matter of computing power and that predictions could be made at 10 m if desired. However, many researchers refrain from using such fine resolutions for global products because these products tend to perform poorly locally and can mislead soil management decisions. The resolution choice should be justified.

Author Response:

While satellite data are now publicly available at spatial resolutions as fine as 10 m, studying forests, land use, and soil properties at that level of detail is often problematic, unnecessary, and inefficient. A 100 m resolution (1 ha) is a more appropriate choice for several reasons:

1. **Ecological and land-use processes operate at coarser scales.**
   Many ecological and land-use processes, including forest composition, land cover transitions, and management zones, typically manifest at coarser spatial scales, generally either at landscape scale or disturbance scales. When considering both natural and anthropogenic disturbance patterns, a minimum mapping unit of 1 ha (100 m) is a reasonable scale for representing vegetation and soil parameters.
2. **Soil properties exhibit meaningful variation at the hectare scale.**
   Soil heterogeneity is often structured around ~1 ha units, whereas finer resolutions tend to capture noise rather than ecologically relevant patterns (McBratney et al., 2003; Delbari et al., 2011).
3. **Aggregating pixels improves prediction accuracy.**
   The predictive power of remote sensing data improves significantly when aggregating pixels. At 10 or 30 m resolution, reflectance measurements are affected by calibration issues, measurement noise, satellite geometry, atmospheric conditions, georeferencing errors of a pixel and more, and small-scale heterogeneity unrelated to surface parameters. Machine learning or analytical models perform better at

aggregated scales with less reflectance errors than at the level of individual noisy pixels. Consequently, most high-resolution maps at 10 or 30 m do not achieve true pixel-level accuracy and are often mistaken for being more precise than they actually are.

It is important to distinguish between map resolution and pixel spacing; posting data at 10 or 30 m does not imply that parameters are accurately resolved at that scale, as pixel spacing for gridding and the true prediction resolution are separate concepts. The spacing of the pixels does not automatically define the scale at which the model's predictions are reliable. Our 100 m maps could be applied to finer pixel grids (10 m or 30 m) using the same model, with the only difference being computational cost. However, a 100 meter resolution provides a more ecologically appropriate and computationally efficient scale for landscape-level analyses.

Reference:

Delbari, M., Afrasiab, P., and Loiskandl, W.: Geostatistical analysis of soil texture fractions on the field scale, Soil and Water Research, 6, 173–189, https://doi.org/10.17221/9/2010-SWR, 2011.

McBratney, A.B., Mendonça Santos, M.L., and Minasny, B.: On digital soil mapping. Geoderma, 117(1–2), 3–52, https://doi.org/10.1016/S0016-7061(03)00223-4, 2003.

Model validation. The authors used model validation based on data splitting. Repeated cross-validation should be used instead. Data splitting can lead to accidentally high or low validation statistics depending on the split, and it might also be susceptible to fraud if one were to select a split that yields better results. It should be replaced.

Author Response:

Thank you. During model development, we conducted K-fold cross-validation with 50 simulations. We added further detail to **Methods section 2.6, Results section 3.3** and provide the results in **Table 1** of the revised manuscript.

We note that all machine learning and regression models can suffer from overfitting, particularly at the tails of predicted distributions, which is amplified when working with noisy data. Climate and remote sensing data, when used to assess SOC, are considered noisy due to the weak statistical correlation between spectral information and SOC values. This can lead to overestimation of low values and underestimation of high values, distorting the mean SOC distribution and introducing systematic errors at the extremes. To mitigate this, we use a stratified approach, training biome-specific models. This allows for more localized SOC

estimation, substantially reducing global overfitting and minimizing biases in the tails of SOC distributions.

Previously, we applied a bias-correction approach (Xu et al., 2017, 2021) to improve accuracy in the distribution tails, analogous to histogram matching. While this bias-correction algorithm significantly improves distribution bias, it does introduce a slight increase in random error at the pixel level. We removed this bias correction, accepting small residual biases (<1% in most regions), as biome-specific modeling already mitigates bias, and the impact of these residual biases on large-scale SOC estimates is minimal.

Bias correction. Applying a post-processing bias correction is, to my knowledge, never seen in digital soil mapping. This is because most models we use (geostatistics, random forest) have little to no bias. Are there even biases in the predictions? This is not common for random forest, and the explanation given by the authors on this aspect does not seem relevant to the work (particularly for users applying the map for soil carbon accounting).

Author Response:

In the previous version of the soil carbon map, a global histogram matching bias correction was applied to the combined outputs of the three broad models (global, mangrove, peatland) to reduce systematic over- or underestimation relative to observed values. In the current version, we modeled soil carbon separately for 14 biomes and ecosystems, with each model trained on a more homogeneous subset of data. This biome-specific approach inherently reduces systematic bias, as each model is better able to capture local ecological patterns. Therefore, no additional bias correction was applied, and model outputs are reported directly from the biome-specific models.

This clarification has been added to **Methods Section 2.6**, as follows:

> "The biome-specific approach inherently reduces systematic bias, allowing each model to better capture local patterns. Consequently, no additional bias correction (i.e. histogram-based bias correction) was applied, and the reported values reflect the direct outputs of the biome-specific models."

Uncertainty assessment. The uncertainty assessment needs to be completely redone. In digital soil mapping, one is interested in a prediction interval, not a confidence interval. This is common knowledge and can be found in standard DSM textbooks and papers. A confidence interval is of limited interest, as it does not inform us about the uncertainty of new observations. The authors should therefore obtain a prediction interval. Second, the idea of bootstrapping a random forest model is not meaningful. Random forest already uses bootstrapping internally, so the authors are effectively bootstrapping twice. A much simpler approach, and one that is widely implemented, is to use quantile regression forest, which directly reports a prediction interval. This is probably the most common procedure in DSM to

generate uncertainty intervals with machine learning. Third, the approach based on Z-scores relies on the assumption of normally distributed errors, which is generally not valid for machine learning.

Author Response: Regarding error reporting, we have enhanced our uncertainty reporting to provide detailed information about model performance. Total propagated uncertainty including model variance and residual variance are now included to support inferences over large areas, although they are not applied to spatial information. Additionally, as the reviewer suggested, we used random forest tools to calculate all associated errors. However, these tools do not provide pixel-based uncertainty; instead, pixel-level uncertainty maps must be calculated using a bootstrapping approach, which involves varying the training and testing data pools. We have not done any double bootstrapping for pixel level confidence interval or prediction intervals.

Our mapping pipeline employs this bootstrapping method for all types of machine learning models. While the random forest (RF) model includes additional tools for using Quantile Regression Forests (QRF) to predict conditional quantiles of the response variable, not just the conditional mean, the end results are often similar, as demonstrated in numerous publications. For this study, we provided 95% confidence intervals for the pixel values of mean SOC.

Residual autocorrelation. Once the above issues are addressed, the authors should check for residual autocorrelation and report the fitted variograms. It may be that kriging of the residuals is needed if autocorrelation remains.

The many major comments above suggest that the maps may present a misleading representation of the spatial patterns and average or total stocks. These points need to be addressed very seriously, and the authors should ensure familiarity with the state of the art in digital soil mapping, as most of these are standard procedures in the field.

Author Response:

We also evaluated residual autocorrelation, an issue that has been addressed in several of our group's previous publications (Weisbin et al., 2014; Saatchi et al., 2011; Xu et al., 2016, 2017, 2021; McRoberts et al., 2021; Cushman et al., 2023), and found that the impact of residual autocorrelation on regional-scale inferences is minimal. Its influence becomes relevant primarily when estimating the uncertainty of SOC means and totals for small regions. At larger spatial scales, however, semi-variograms indicate a rapid exponential decline in autocorrelation with distance (Weisbin et al., 2014; Xu et al., 2017).

References:

Cushman, K. C., Albert, L. P., Norby, R. J., and Saatchi, S.: Innovations in plant science from integrative remote sensing research: an introduction to a Virtual Issue, New Phytologist, 240, 1707–1711, https://doi.org/10.1111/nph.19237, 2023.

McRoberts, R. E., Næsset, E., Saatchi, S., and Quegan, S.: Statistically rigorous, model-based inferences from maps, Remote Sensing of Environment, 279, 113028, https://doi.org/10.1016/j.rse.2022.113028, 2022.

Saatchi, S. S., Harris, N. L., Brown, S., Lefsky, M., Mitchard, E. T. A., Salas, W., Zutta, B. R., Buermann, W., Lewis, S. L., Hagen, S., Petrova, S., White, L., Silman, M., and Morel, A.: Benchmark map of forest carbon stocks in tropical regions across three continents, Proceedings of the National Academy of Sciences, 108, 9899–9904, https://doi.org/10.1073/pnas.1019576108, 2011.

Weisbin, C. R., Lincoln, W., and Saatchi, S.: A Systems Engineering Approach to Estimating Uncertainty in Above-Ground Biomass (AGB) Derived from Remote-Sensing Data, Systems Engineering, 17, 361–373, https://doi.org/10.1002/sys.21275, 2014.

Xu, L., Saatchi, S. S., Yang, Y., Yu, Y., and White, L.: Performance of non-parametric algorithms for spatial mapping of tropical forest structure, Carbon Balance and Management, 11, 18, https://doi.org/10.1186/s13021-016-0062-9, 2016.

L. Xu, S. S. Saatchi, A. Shapiro, V. Meyer, A. Ferraz, Y. Yang, J.-F. Bastin, N. Banks, P. Boeckx, H. Verbeeck, S. L. Lewis, E. T. Muanza, E. Bongwele, F. Kayembe, D. Mbenza, L. Kalau, F. Mukendi, F. Ilunga, D. Ebuta, Spatial distribution of carbon stored in forests of the Democratic Republic of Congo. *Sci. Rep.* **7**, 15030. 2017.

Xu, L., Saatchi, S. S., Yang, Y., Yu, Y., Pongratz, J., Bloom, A. A., Bowman, K., Worden, J., Liu, J., Yin, Y., Domke, G., McRoberts, R. E., Woodall, C., Nabuurs, G.-J., de-Miguel, S., Keller, M., Harris, N., Maxwell, S., and Schimel, D.: Changes in global terrestrial live biomass over the 21st century, Science Advances, 7, eabe9829, https://doi.org/10.1126/sciadv.abe9829, 2021.