Response to Reviewer 3

The manuscript titled "Mapping global distributions, environmental controls, and uncertainties of apparent top- and subsoil organic carbon turnover times" addresses an important knowledge gap and is therefore likely to be of broad interest. The primary aim of the study is to map carbon turnover times for the topsoil (0–30 cm) and subsoil (30–100 cm) at a global scale, with a spatial resolution of 30 arcsec. The resulting maps, along with their associated prediction uncertainties, are freely available on Zenodo.

Overall, I am satisfied with the manuscript and the dataset provided on Zenodo. However, I recommend a few minor revisions before it is suitable for publication. These are as follows:

AC: Thank you for your positive general comment. Our responses to your specific comments as below.

(1) A few minor spelling and grammatical corrections are needed to enhance the text's readability.

AC: Thank you for pointing this out. We have revised the manuscript to correct these errors in language.

(2) I am curious why the bias (mean error) was not calculated. Demonstrating that the maps are unbiased would strengthen the validity of the results.

AC: We appreciate the reviewer's suggestion regarding the calculation of bias (mean error) in addition to RMSE. In our study, we primarily used RMSE (root mean squared error) as the validation metric, as it accounts for both the variance of errors and their magnitude, making it a robust measure of prediction accuracy. However, we acknowledge that assessing bias (mean error) (here we think the metric name should be MAE [mean absolute error]) would provide additional insight into systematic over- or under-predictions. To address this, we have now calculated the mean error (ME) alongside RMSE and included these results in the supplementary material (Table S3, also show as below). The MAE values confirm that the predicted τ values do not exhibit significant systematic bias, further supporting the validity of our mapping approach.

Table. Global cross-validation results for top- and subsoil τ predictions.

Variable to predict	RMSE	MAE	R^2	MEC
Topsoil τ	14.60	5.72	0.87	0.93
Subsoil τ	175.05	65.57	0.70	0.83

Note: RMSE, root mean squared error; MAE, mean absolute error; R^2, coefficient of determination; MEC, model efficiency coefficient.

(3) I am also wondering why the R-squared value was used instead of the model efficiency coefficient (MEC) (https://en.wikipedia.org/wiki/Nash%E2%80%93Sutcliffe_model_efficiency_coefficient). MEC is often considered more appropriate than R-squared in similar contexts.

AC: We appreciate the reviewer's suggestion regarding the use of the model efficiency coefficient

(MEC) as an alternative to R^2. In our study, we chose R^2 as a measure of model performance because it is widely used in machine learning-based predictive mapping to assess the proportion of variance explained by the model. Additionally, R^2 allows for direct comparison with previous studies that have employed similar modeling approaches. We acknowledge that MEC is also particularly useful in environmental modelling and soil mapping contexts. Thus, to further strengthen our validation, we have now computed MEC alongside R^2, and included these results in the supplementary material (Table S3). The MEC values are generally consistent with R^2, reinforcing the reliability of our predictions.

(4) Clarification of Section 2.4.2 (Quantification of uncertainty): I believe this section could be revised to provide clearer and more reproducible methodological details. If I understand correctly, did you fit a single QRF model using approximately 90,000 × 100 carbon turnover observations to account for input uncertainties? Or did you fit 100 different QRF models and then combine the results? Clarifying this would enhance transparency.

AC: In our study, we considered the potential error (represented by the calculated standard deviation [SD] input variables) in estimating τ for each sample point. Then, a Monte Carlo approach was adopted to incorporate the SD in the estimated τ at each sample. That is, the value of τ at each sample location was randomly drawn 100 times from a normal distribution given the known mean and SD. Therefore, we generated 95,200 × 100 and 66,807 × 100 samples for topsoil and subsoil, respectively. Finally, we used these samples to fit a QRF model, and the model can directly produce the 0.05 quantile and 0.95 quantile of predictions so that helped us obtain uncertainty maps. We revised some sentences in Section 2.4.2 to make this description clearer.

(5) It would be beneficial to include guidance on how to interpret the PIR values. This would help readers better assess the quality of the maps and understand the magnitude of the prediction uncertainty.

AC: Yes, we agree to improve the explanation of prediction interval ratio (PIR) in the manuscript to help readers have a better understanding. PIR provides a relative measure of uncertainty by comparing the width of the prediction interval to the predicted median value. A lower PIR indicates higher confidence in the predictions, while a higher PIR suggests greater uncertainty. To improve clarity, we added the guidance on how to interpret PIR values in the revised manuscript.

(6) Similarly, a more detailed explanation of PICP would be helpful. Many readers may be unfamiliar with this metric, including what it measures, how to interpret it, and how it reflects the fairness of the uncertainty quantifications.

AC: We agree to add explanation of prediction interval coverage probability (PICP) would be helpful, as many readers might be unfamiliar with it. PICP is a metric used for evaluating the reliability of uncertainty quantifications by measuring the proportion of observed values that fall within a given prediction interval. It can effectively evaluate if the probability assigned to the prediction intervals is equal to the frequency of empirical test data within the prediction intervals (Goovaerts, 2001; Malone et al., 2011; Schmidinger and Heuvelink, 2023) To improve clarity, we have expanded the explanation of PICP in the revised manuscript. We clarify that: PICP is defined as the fraction of observed values that fall within the predicted uncertainty bounds (e.g., the 90% prediction interval). A PICP of 0.90 indicates that 90% of observed values lie within the computed prediction interval, suggesting well estimated uncertainty. If PICP is significantly lower than the expected interval probability (e.g., < 0.90 for a 90% interval), it implies that the model underestimates uncertainty, potentially leading to overconfidence in predictions. Conversely, if PICP is much higher than the expected probability (e.g., > 0.90 for a 90% interval), it suggests overestimated uncertainty, meaning the prediction intervals may be too wide. We also put more detailed explanation into the caption of Figure S22 in supplementary material.

References:

Goovaerts, P.: Geostatistical modelling of uncertainty in soil science, Geoderma, 103, 3–26, https://doi.org/10.1016/S0016-7061(01)00067-2, 2001.

Malone, B. P., McBratney, A. B., and Minasny, B.: Empirical estimates of uncertainty for mapping continuous depth functions of soil attributes, Geoderma, 160, 614–626, https://doi.org/10.1016/j.geoderma.2010.11.013, 2011.

Schmidinger, J. and Heuvelink, G. B. M.: Validation of uncertainty predictions in digital soil mapping, Geoderma, 437, 116585, https://doi.org/10.1016/j.geoderma.2023.116585, 2023.

(7) The sections on mapping and uncertainty quantification appear somewhat intertwined. Did you first fit an RF model to create the maps and then use a separate QRF model (including the inputs' uncertainty) to quantify the uncertainty associated with the RF predictions? Did you fix the (pseudo-)random number generator to maintain consistency between the RF and QRF models? Clarifying these aspects would enhance the transparency and reproducibility of the methodology.

AC: We first trained a standard RF model to generate the spatially explicit predictions of τ. Then, a QRF model was adopted for uncertainty quantification. To maintain consistency between the RF and QRF models, the QRF model was trained using the datasets. We also fixed the pseudo-random number generator during the training procedure for these two models. This ensures that both models operated with the same data splits and bootstrapped samples, minimizing variations introduced by randomness and enhancing reproducibility. To improve clarity, we have added this description in Section 2.4.2.

Once these minor issues are addressed, I believe the manuscript will be ready for publication.

AC: We thanks the reviewer for this valuable suggestion above, which has helped us enhance the quality of our manuscript.