Response to the Reviewers

Reviewer: 1

Q: Accepted subject to technical corrections.

Re: We highly appreciate your time again.

Reviewer: 2

Q: Thank you for the revision. However, there is a huge bias in your machine learning predictions which might cause damage to your study and conclusions. The machine learning algorithm has large bias based on figure S6b. The slope is far from 1. Looks like a 0.5. This is huge bias, even though R^2 is ok.

Re: Thank you for your suggestion. We have replaced and added some environmental factors according to reviewer 3's suggestion. Meantime, we have given the slope and R^2, and the results are pretty good. At the same time, we have shown the corresponding slope of RF in different ecosystems. Thanks again.

Reviewer: 3

Q: This study synthesizes global measurement of DOC data and the corresponding environmental variables to develop a random forest model. This ML model is applied to interpret key environmental regulators and deliver a global mapping of DOC concentration. The development of such an ML model can benefit the quantification of DOC concentration at the global scale and the assessment of soil carbon stability under climate change. However, the manuscript is not written clearly and needs a serious and major revision to make it publishable. (Please note that all line numbers I cited here are based on the tracked version of the manuscript).

Re: Thank you for your recognition of the innovation of our manuscript and your good suggestions and comments. Next, we will reply to your suggestions and comments one by one.

Q 1: First, the research method and its description need to be refined. Please see my specific comments below.

(1) using “month” as an input feature of the ML model for predicting DOC concentration at the global scale is not reasonable. Hydrothermal patterns during the same month can be highly different in the northern hemisphere and the southern hemisphere (e.g., summer climate vs. winter climate). Because of this difference, applying “month” to cluster the DOC concentration is usually binary, which may underestimate its regulation on the DOC concentration. Alternatively, the direct application of monthly

properties (e.g., hydrothermal conditions) as an input feature will make more sense.

Re: Thank you for your suggestion. We have used the "monthly evaporation, seasonal variability of precipitation, and seasonal variability of temperature" index instead of "month" (Fig. 2). The indicators of monthly temperature and precipitation were not used because their interpretation rates were low. The corresponding content has been further described in the Materials and Method.

(2) On lines 165-166: Although soil C is strongly correlated to soil N, the soil C/N ratio can be a good indicator of soil nutrient availability, which will affect the decomposition of soil organic matter and soil DOC concentration. Therefore, Fig S3 is not a reasonable justification for excluding soil N from the input features. In fact, the more reasonable approach could incorporate soil C/N ratio into the input feature and then determine if it should be excluded based on the Gini value analysis of important environmental factors.

Re: Done. The indicator of soil C/N ratio has been used as the predictor. IncNodePurity was used to exclude some factors, such as ecosystems (Fig. S6).

(3) On lines 129-142: Soil data is highly heterogeneous, and soil DOC concentration can be highly sensitive to heterogeneous soil environments. It's unclear if the collection of environmental data (especially soil texture data) corresponding to the 12807 DOC observations is sample-level or site-level. According to the reported number of these environmental variables (e.g., soil texture, pH, MBC, etc.) in Table 1, I believe they are site-level. In addition, the author has mentioned that missing data for some sites are extracted from gridded soils. Table S2 clearly shows that there is a significant bias between measured and extracted soil variables (e.g., bulk density, soil texture, and MBC). These scale inconsistencies between DOC concentration data and the corresponding environmental factors need to be clearly described in the method section. A clear description of the size of environmental data at each level will be helpful for quantifying uncertainty from linking difference scales of environmental variables with collected DOC concentration data.

Re: Thank you for your comments and suggestions. For the first question, the 12807 DOC observations is sample-level. It has been revised as "Based on these criteria, we compiled a total of 12,807 DOC observations based on 1610 sites from 975 publications". For the second question, there was a significant bias between measured and extracted soil variables (Fig. S2). We have added a description of the resolution and the result. It has been revised as "We extracted elevation, MAT, MAP, monthly evaporation (ETM), seasonal variability of precipitation (SVP), and seasonal variability of temperature

(SVT) data from WorldClim Version 2 (https://www.worldclim.com/) with resolution of 1 km × 1 km, ecosystem data from NASA's Socioeconomic Data and Applications Center (https://sedac.ciesin.columbia.edu) with resolution of 1 km × 1 km, soil properties from OpenLandMap version 2.0.0 (https://openlandmap.org) with resolution of 0.25 km × 0.25 km, and microbial biomass carbon data from the open database of figshare (https://doi.org/10.6084/m9.figshare.19556419) with resolution of 1 km × 1 km. Despite bias, there is a significant linear relationship between the measured values and the corresponding extracted values (Fig. S2). Noteworthy, this bias could introduce some uncertainty to the results".

(4) On lines 166-167: It's unclear what variables are excluded due to the lack of data.

Re: It has been revised as "Further, we did not include some variables (e.g., soil moisture, soil porosity, ferroaluminum oxide, microbial structures, microbial diversity, and carbon cycling enzymes) because they were rarely report in the target papers".

(5) Figure S8 showed the sensitivity analysis of applied environmental features in the RF model, but no description of this analysis method in the method section.

Re: Done. It has been revised as "For evaluate the sensitivity analysis of model predictions, the Sobol index, a variance of based global sensitivity analysis method, was used to assesses how model input parameters impact output results (Fig. S9). It breaks down the system's total variance into contributions from individual inputs and their combinations". Meantime, we have added the description of interaction effects between key drivers of derived soil DOC concentration (Fig. 4) as "For explore the interaction effects between key drivers of derived soil DOC concentration, SHapley Additive exPlanations (SHAP) is used to interpret machine learning model predictions by calculating the contribution of features to the model's predictions (Fig. 4). SHAP values can be further decomposed into main effects and interaction effects, where interaction effects reveal the interactions between features. SHAP interaction values are obtained by first defining an explainer using the TreeExplainer function (by passing the model to it), and then deriving the interaction values from this explainer. These values can be interpreted similarly to standard SHAP values, explicitly quantifying how individual features and their pairwise interactions contribute to specific predictions".

Q 2: The description and discussion of results are not clear and have some logical errors.

(1) On lines 287 -294: The description of importance feature analysis is not consistent with Figure 2a. First, the statement "soil properties were the most important predictor categories" is not accurate.

Except for MAT, MAP, and Ecosystem, almost all other features are soil properties. Figure 2 only shows that elevation plays the most important role, followed by soil clay fraction. Then SOC, MAT, and MAP almost play the same role. Second, "their contributions were lower than those of the top four predictors" on line 294 is inconsistent with the figure 2a, there is even no four predictors above these features discussed here. Third, the RF model required the input features to be relatively independent, although, to some extent, correlation among input features is unavoidable. The author mentioned that the elevation strongly correlated with soil pH, bulk density, and microbial biomass carbon. If this is the case, you may consider excluding correlated environmental factors and retrain/test the RF model and see if the model could still have a similar performance. In summary, the discussion of Figure 2a is not consistent with the figure itself and needs a major revision.

Re: For the first question, it has been revised as "Elevation played the most important predictor for soil DOC prediction among the selected 14 variables, followed by SOC, SVT, and soil clay". For the second question, it has been revised as "The relative importance of MAP, SVP, MBC, soil pH, soil sand, and soil C:N was gradually diminishing". For the third question, we have tried to exclude a number of influencing factors, but the accuracy of the prediction results has decreased significantly, except for excluding ecosystems (Fig. S6). Therefore, we have retained these 14 factors as drivers.

(2) On lines 295-299: Partial dependence analysis is a visualization way to interpret the RF model. However, the accuracy of PD calculation strongly determines the feature distribution. PD calculation usually overinterprets regions with limited/no data. Therefore, figure 3 should be improved by overlaying the feature distribution histogram. According to the MAT distribution shown in Figure S4, the description of negative soil DOC with MAT needs to be careful, as MAT less than -5 degrees has limited data. In addition, PD analysis requires the independent of features. As elevation is strongly correlated with other features, using PD to interpret elevation relationship with DOC also needs to discuss the uncertainty.

Re: Thanks for your good suggestions. For the first question, Fig. 3 has been improved by overlaying the feature distribution histogram. You are right that it is necessary to be cautious in describing the relationship between PD and factors under the limited data. This is why we present Fig. S4. At the same time, we have also revised some of the descriptions to better reflect the real situation as "We found a positive correlation between soil DOC and both elevation and soil organic carbon, although there were fewer data points corresponding to higher elevations and greater soil organic carbon values

(Fig. 3f). Soil DOC showed a trend of decreasing first and then increasing with the increase of MAT (0-30 ℃), SVT (0-1.5), and soil clay (0-50%) (Fig. 3a, d and h). Soil DOC showed a trend of decreasing first and then stabilizing with the increase of soil depth and soil pH (4-8.5). The inflection point of soil depth and soil pH was 10 cm and 5.8, respectively (Fig. 3i and k)". We have also added a description of the interaction as "Elevation, SOC, SVT, and soil clay had strong negative interactions with MAT (Fig. 4). This means as the MAT variable increases, the influence of the other variables is weakened. Elevation had a positive interaction with bulk density, suggesting they work together to affect soil DOC".

(3) On lines 331-343: the comparison of this study's results of DOC concentration in the forest and grassland with other studies (e.g., Cai et al., 2021, Perrot et al., 2023, Deng et al., 2023) isn't in a clear way. First, unlike SOC, DOC concentration can fluctuate across time and space, it's unclear if these studies reported DOC concentration at similar time in the same location. Without providing this information, the comparison doesn't make sense. Second, different types of grassland ecosystems and forest ecosystems (e.g., tropical forest, boreal forest) can be very different in terms of decomposition ability, the citation of "the cooler conditions in forest soils limits microbial activity and organic matter decomposition, reducing DOC concentration" could not justify whey tropical forest region also has showed lower DOC concentration than the grassland region. I believe the mineral adsorption effect may be a reason, but have not been discussed carefully here. Considering the high variation in DOC concentration in space and time, I think this manuscript should focus on interpreting the main regulator of DOC concentration and its possible space differences instead of mapping DOC concentration. The model itself will be more useful for modeling SOM decomposition than the provided DOC mapping. Therefore, Section 4.2 discussion should be the main delivery of this manuscript.

Re: We sincerely appreciate the reviewer's insightful critiques. We have focused on the drivers. The section 4.1 (ecosystem comparison) has been removed. At present, section 4.1 is divided into elevation and soil properties. The section 4.2 focuses on the influence of SVT in climate factors on soil DOC. Due to the excessive content of the revisions, detailed revisions can be found in the Discussion section of the manuscript.

(4) Section 4.3: The interpretation of section 4.1 and section 4.3 looks a little repeated. I understand that section 4.1 focuses on interpreting collected datasets, section 4.3 focusses on interpreting the mean DOC pattern generated in this study. Considering the high variability of DOC concentration in space

and time, this interpretation for section 4.3 will make more sense compared with section 4.1.

Re: Thanks for your comments. Section 4.1 focuses on interpreting collected datasets under different ecosystems and section 4.3 focuses on interpreting the mean DOC pattern based on RF model. We agreed with your view. Therefore, we have deleted the section 4.1. Thanks again.

Q 3: Finally, the manuscript is not written clearly and still has several unclear descriptions and typos. Please see my specific comments on a few examples as follows.

Re: Thank you for your good suggestions and comments. We have replied one by one.

Q 4: On line 162 and line 387: elevation is not a climate factor.

Re: Yes, elevation has been grouped into a separate category.

Q 5: On lines 230-232: Input data listed in Table S1 has different spatial resolution. A clear description of how these data are resampled to the unified spatial resolution needs to be provided. In addition, figure S5 seems to be not relevant to the sentence that cited this figure here. There may be some typos.

Re: For the first question, we have added the description as "Duo to the different spatial resolution of input variables data, resampling techniques enables the conversion of raster data between spatial resolutions to facilitate spatial analysis and modeling. The core principle of resampling involves estimating pixel values at new resolutions through interpolation or other mathematical methods. Specifically, down-sampling (high-to-low resolution conversion) requires aggregating values from multiple high-resolution pixels into a single low-resolution pixel. Up-sampling (low-to-high resolution conversion) necessitates generating new pixel values through interpolation algorithms". For the second question, figure S5 has been changed to figure S11 (the relative uncertainties of global predictions).

Q 6: On line 249 and line 251: Figure 2 cannot reflect the text that cited this figure. I believe you plan to cite Figure 3.

Re: It has been modified to Figure 3.

Q 7: On line 284: Figure S3 is not relevant to the sentence that cited this figure.

Re: It has been modified to Table S2.

Q 8: On lines 284-285: The sentence is not clearly written and has grammar issues.

Re: It has been revised as "The relative importance of soil DOC drivers and the global map of soil DOC distribution were derived from the RF model outputs".

Q 9: On line 286, Table S2 shows that the R2 of the RF model is 0.70, which does not consist of the text description here.

Re: I am very sorry that it was not modified during the first modification. According to the latest results, the R^2 of the RF model has been replaced at 63%.

Q 10: On line 290: add “carbon” after “soil organic carbon”

Re: Done.

Q 11: On line 384: change “After to” to “After”.

Re: Done.

Q 12: On line 361: delete “controlled”. Also this subtitle is inconsistent with the content of this section, as this section didn’t discuss anything about the climate effect.

Re: Done. It has been revised as “Effects of elevation and soil properties on soil DOC concentrations”.