Response to the Reviewers

Reviewer: 1

Q: In this study, the authors produced a global map for soil dissolved organic carbon (DOC) concentration and attributed the global variation to environmental factors. It is an interesting research endeavor, and relevant to the scope of this journal. However, I have a few major concerns.

Re: We highly appreciate your time again. We have now substantially improved the manuscript.

Q1: DOC can fluctuate significantly over time. For example, one may see high concentration in growing season and low concentration in non growing season. As a result, the global map produced in this study may contain considerable uncertainties due to the influence of temporal variation.

Re: Thanks for good comment and suggestion. Indeed, DOC concentrations fluctuate significantly between seasons, typically higher during the growing season and lower during the non-growing season. Therefore, we designed the framework to consider the effects of seasonal changes on the DOC concentration. For the manuscript, sampling time (month) is used to represent the influence of seasonal changes on DOC concentration. We have added a detailed description in the part of 2.4 Global soil DOC mapping:

Soil DOC concentration varied significantly with temporal changes and soil depth (Fig. 2). Sampling time (month) was used to represent seasonal variations in soil DOC concentration. Soil DOC concentration decreased with soil depth and reached a turning point at approximately 10 cm (Fig. 2). Therefore, when extrapolating the RF model to the entire globe, we used a month range from 1 to 12 and depths of 5 (0–10 cm) and 20 (10–30 cm). From this, we generated a total of 24 maps of global soil DOC concentration. We combined these 24 maps into a single map representing the global distribution of soil DOC concentration based on soil depth. Finally, we calculated the global soil DOC stock using the following equation applied to the combined map of global soil DOC concentration:

$$SOC_s = \sum SOC_i \times BD_i \times (1 - f) \times T \times M_i \quad (2)$$

where SOC_s is SOC stock and SOC_i is SOC concentration. The subscript i is the number of global grid. BD, f, and T are soil bulk density, the volumetric percentage of coarse fraction (>2 mm), and the depth of soil layer, respectively. M is the effective area of each grid.

Q2: Methods, especially global mapping and calculation are too simplified and even missing. Data leakage due to training and validation may inflate the RF model performance. (See details below).

Re: Thanks for good comment and suggestion. We have recalculated and reproduced a global

distribution map of soil DOC. In every instance, the models were evaluated using Monte Carlo cross-validation with 100 iterations, employing a 70/15/15 split between training, validation, and testing sets. The root mean square error and R^2 values were calculated to evaluate model accuracy and residual variance, which served as criteria for ranking model performance. A 10-fold cross-validation method was used to evaluate model performance. We have added the validation and test of random forest models for soil DOC (Fig. S6) and the residual plot the model train, validation, and test (Fig. S7).

Q3: The language requires significant improvement.

Re: The grammatical expression has been revised by a native speaker.

AMERICAN
ENGLISH EDITING

EDITING CERTIFICATE

Date: December 12, 2024

Manuscript Authors: Tianjing Ren and Andong Cai

Manuscript Title: *Global patterns and drivers of soil dissolved organic carbon concentrations*

To whom it may concern,

I had the pleasure to perform English editing on the manuscript titled above. I am a native English speaker and hold a doctoral degree in the biological sciences.

I am confident that the English language in the manuscript is of a sufficient standard to warrant a consideration for peer review and/or publication. If needed, we look forward to making revisions as requested by the editor and/or reviewers.

Sincerely,

Russell Doughty, PhD

Russell Doughty, PhD
Research Scientist
Norman, Oklahoma, USA
russ@americanenglishediting.com
https://www.researchgate.net/profile/Russell-Doughty

Q4: L123: Elevation does not belong to climate, but topography.

Re: Done.

Q5: L128-153: While testing a model, please split the data into training, validation, and testing datasets. Data leakage exists without testing datasets. Therefore, your results may be over-confident.

Re: We have modified it in the Material method. See Q2 for detailed responses.

Q6: L147-153: move this part to results section for model performance evaluation. Please also expand it to include more information.

Re: This part has been moved to the results section. We have also expanded this section.

Q7: L188-190: not clear what the authors are trying to say

Re: It has been revised as "The corresponding map of relative uncertainty of prediction was built by displaying the standard deviation divided by the mean prediction, based on our final random forest RF model. The standard deviation reflected the range of possible predictions derived from the iterative build-up of decision trees after 500 model runs".

Q8: L195: it is not normal distribution because the values are natural logged.

Re: It has been revised as "the natural logarithm of soil DOC conformed to a normal distribution".

Q9: L198-200: please provide median in addition to mean.

Re: Done.

Q10: L202-212: RF model performance may be inflated due to data leakage during model training. The authors need to set aside a testing dataset in addition to training and validation datasets.

Re: Done. See Q2 for detailed responses.

Q11: L214-226: Month and depth are predictors in the RF model. What values did the "Month" and "depth" take while extrapolating RF to the whole globe? Please also use medians instead of mean to describe DOC concentration due to its strong skew from normal distribution. How did the authors derive DOC in Pg from mg/kg?

Re: See Q1 for detailed responses. For the first question, sampling time (month) was used to represent seasonal variations in soil DOC concentration. Soil DOC concentration decreased with soil depth and reached a turning point at approximately 10 cm (Fig. 2). Therefore, when extrapolating the RF model to the entire globe, we used a month range from 1 to 12 and depths of 5 (0–10 cm) and 20 (10–30 cm). For the second question, we have used medians instead of mean to describe DOC concentration. For the third question, the specific function has been added in the Materials method.

Q12: L232: please compare median.

Re: Done.

Reviewer: 2

Q: The study aims to understand the global distribution and driving factors of soil dissolved organic carbon (DOC) concentrations. Objectives include Identifying global patterns of soil DOC concentrations, determining primary factors controlling soil DOC concentrations globally, and quantifying global soil DOC storage. This work is highly appropriate for Earth System Science Data as it presents a valuable new global dataset with clear methodology and thorough documentation.

Re: Thank you for your high evaluation of our research work.

Q1: Novel contribution: A comprehensive global soil DOC database to date (12,807 observations from 975 publications). Application of machine learning to identify key drivers and predict global patterns. More accurate estimation of global soil DOC stocks compared to previous studies.

Re: Thank you very much for your recognition and valuable comments on our research work. The few contributions you pointed out are really the highlights of our research and we are very proud of them.

Abstract

Q2: Add brief mention of validation metrics for the machine learning model

Re: Done. It has been revised as “Machine learning techniques were employed, including 10-fold cross-validation and evaluating model performance by R-squared and root-mean-square error, to predict the relative importance of various predictors and the global distribution of soil DOC concentrations”.

Q3: Include the temporal range of the compiled data

Re: Done. It has been revised as “Here, we compile a comprehensive global database of soil DOC concentrations, encompassing 12,807 observations extracted from 975 scientific publications published between 1984 and 2020”.

Q4: Specify the spatial resolution of the global predictions

Re: It has been revised as “Using these findings, a global map of predicted soil DOC concentrations was produced at a 0.05 °by 0.05 °resolution”.

Q5: L19: Specify the time period over which these samples were collected

Re: Done. See Q3 for detailed responses.

Q6: L21: After "Machine learning techniques", specify which ones were used

Re: Done. See Q2 for detailed responses.

Introduction

Q7: Consider adding a brief discussion of temporal variations in DOC

Re: We have added the temporal variations in DOC.

Some studies have reported large temporal variations in soil DOC concentrations at certain field sites (Ding et al., 2022; Zhao et al., 2022), with significantly higher DOC concentrations in summer and autumn than in winter and spring. Seasonal effects on soil DOC concentrations are closely associated with factors such as precipitation, soil moisture, and substrate availability (Ren et al., 2023). In warmer seasons, soil DOC production can increase due to active organic matter decomposition, driven by higher microbial activity, as well as greater DOC contributions from root exudation during periods of more active plant photosynthesis.

Q8: Include more recent references (post-2020) on global carbon cycling

Re: Done as your suggestion.

Q9: Expand on the limitations of previous global DOC mapping efforts

Re: It has been revised as “However, these maps are subject to considerable uncertainties due to limited data, restrictive factor selections, and low interpretation rates. First, the global soil DOC maps produced by Guo et al. (2020) and Langeveld et al. (2020) rely on relatively few observational data points (2890 and 762 pairs, respectively). There is a lack of valid observational data for Africa, South America, Eastern Europe, and Central Asia. Second, when assessing the global distribution of soil DOC concentrations, Guo et al. (2020) and Langeveld et al. (2020) have not considered the impact of seasonal changes, even though soil DOC concentrations can vary substantially with shift of season. Third, topsoil DOC concentrations were treated as constant value by Guo et al. (2020) and Langeveld et al. (2020), overlooking the dynamic nature of soil DOC, which decrease with increasing depth. In reality, soil DOC concentrations are higher in surface soils (0-10 cm) and decline with depth, exhibiting a clear vertical gradient. Finally, Guo et al. (2020) and Langeveld et al. (2020) have explained only about one-third of the variation in soil DOC by using multivariate linear equations. Recent advancements in machine learning has enabled researchers to apply such techniques because of their capacities to automate feature extraction, handle large datasets, and identify complex patterns, ultimately offering significant advantages in predictive accuracy and adaptive learning”.

Q10: L36: Update IPCC citation to most recent report

Re: Done.

Q11: L54-55: Add recent examples of "extensive research"

Re: Done. We have given the latest references.

Methods

Data quality:

Q12: Need clearer criteria for handling outliers

Re: We have added a description of how outliers are handled:

Before extracting the target data, we employed the Isolation Forest method for anomaly detection. The algorithm constructs random binary trees, where anomalies are typically isolated more rapidly, while normal points require more splitting steps.

Q13: Should explain how temporal variations were addressed

Re: Thanks for your good suggestion. Soil DOC concentration can fluctuate significantly over temporal variations. We're replacing seasonal variations with sampling times. In the part of 2.3 predictive modeling, the observation time has been replaced with temporal variations:

Soil DOC concentration varied significantly with temporal changes and soil depth (Fig. 2). Sampling time (month) was used to represent seasonal variations in soil DOC concentration. Soil DOC concentration decreased with soil depth and reached a turning point at approximately 10 cm (Fig. 2). Therefore, when extrapolating the RF model to the entire globe, we used a month range from 1 to 12 and depths of 5 (0–10 cm) and 20 (10–30 cm). From this, we generated a total of 24 maps of global soil DOC concentration. We combined these 24 maps into a single map representing the global distribution of soil DOC concentration based on soil depth.

Q14: More detail needed on handling missing predictor variables

Re: Done. It has been revised as “When environmental factors were not reported in original publication, the missing data were extracted from grid datasets according to the geographic coordinates of each observed site (Table S1). We extracted elevation, MAT, and MAP data from WorldClim Version 2 (https://www.worldclim.com/), biome data from NASA's Socioeconomic Data and Applications Center (https://sedac.ciesin.columbia.edu), soil properties from OpenLandMap version 2.0.0 (https://openlandmap.org), and microbial biomass carbon data from the open database of figshare (https://doi.org/10.6084/m9.figshare.19556419)”.

Methodological rigorous:

Q15: Consider using ensemble methods beyond Random Forest

Re: Thanks for your suggestion. Seven ensemble methods were applied to explore the drivers of soil DOC concentrations (Table S2). Random Forest model were identified to estimate the drivers of soil DOC concentrations based on the R^2 and root-mean-square error.

Q16: Add cross-validation across different ecosystems

Re: We have added the cross-validation based on main ecosystems (Fig. S10). Overall, the simulation of each ecosystem was satisfactory.

Q17: Include uncertainty analysis for data standardization process

Re: We have added the uncertainty analysis for data standardization process in the part of 4.4 Limitations and predictive uncertainties:

Fourth, although data standardization enables consistent comparison and analysis of soil DOC across different measurement methods, there were potential issues such as the possible loss of original data characteristics, dependence on accurate parameters, overgeneralization, increasing the complexity of data interpretation, and introducing bias.

Q18: L91: Add justification for each inclusion criterion

Re: It has been revised as “First, we included only data from terrestrial ecosystems (excluding oceans and rivers) to maintain consistency in environmental factors and ecological interactions. Second, we used only topsoil data (0-30 cm) to ensure data representativeness and quantity. Third, we recorded duplicate results from different articles only once to avoid overrepresentation of certain research groups or locations. Finally, we included agricultural soils affected by human activities such as tilling and fertilization but excluded industrial and urban soils to avoid complexity introduced by industrial and urban settings”.

Q19: L102: Explain how missing data quality was assessed

Re: We have added a graph to validate the predictor values extracted from global maps in the attachments (Fig. S2). Although there is a significant linear relationship between the measured values and the corresponding extracted values, especially climate, soil texture, and soil pH, we also emphasized that there was a deficiency in some predictive variables; although we had extracted missing data through gridded datasets, this inevitably introduced uncertainty in predictions, particularly for soil variables. Thanks for your good suggestion, again.

Q20: L140-141: Specify cross-validation fold number

Re: A 10-fold cross-validation method was used to evaluate the performance of the models.

Q21: L150-169: Consider adding a flowchart for model selection process

Re: Done (Fig. S5).

Q22: L170-174: Clarify uncertainty calculation method

Re: We have added the uncertainty calculation method:

When we generated partial dependence with RF, several uncertainties arose. The high model complexity sometimes slowed predictions, especially with many trees. The limited interpretability of the RF models could complicate understanding partial dependence. Sensitivity to noise potentially led to overfitting and reduced accuracy. Variable importance measurements could also be biased by varying feature scales or categories, potentially skewing interpretations of feature-outcome relationships.

Results

Data presentation:

Q23: Include residual plots for model validation

Re: Done (Fig. S7). A residual plot has been placed in the attachment.

Q24: Provide more detailed ecosystem-specific analyses

Re: Done. We have added descriptions of soil DOC concentrations for specific ecosystems (Fig. S10).

Additional analyses

Q25: Interaction effects between key drivers

Re: Done (Fig. S9).

Q26: Sensitivity analysis of model predictions

Re: Done (Fig. S8).

Discussion

Q27: Expand on implications for carbon cycling models

Re: We have added a part of 4.4 Implications for carbon cycling models:

Carbon cycling models are key tools for predicting how soil organic carbon responds to future global changes. Considerable uncertainty exists in simulating and predicting soil organic carbon cycles in many current Earth system models, largely due to model structure, model parameters, and initial conditions (Luo et al., 2015). Regarding model structure, the soil carbon pools in models cannot be

directly separated through experiments, which hamper the quantification of many parameters (Bailey et al., 2018). By integrating global soil DOC concentration data and coupling it with particulate organic carbon, mineral-associated organic carbon, and microbial biomass carbon pools, future models can establish a quantifiable structure based on measurable pools. Our study reveals key factors affecting soil DOC concentrations, such as elevation, soil clay content, and soil organic carbon, can be incorporated into carbon cycle models to improve their predictive capabilities. Moreover, this research provides a detailed global distribution map of soil DOC, which is essential for model parameterization and validation, particularly in regions where data are scarce.

Q28: Discuss potential impacts of climate change on DOC patterns

Re: Done. We have added the effect climate change on soil DOC:

In high-latitude regions, low temperatures limit microbial activity, which slows the decomposition of organic matter and leads to more organic carbon being retained in dissolved form (Patoine et al., 2022) , thereby increasing soil DOC concentrations. In addition, soils in high-latitude areas are often moist or frozen due to low temperatures, limiting oxygen supply and further inhibiting microbial decomposition (Zhou et al., 2024b). These moist or frozen conditions also help protect organic matter, reducing its decomposition and contributing to DOC accumulation. Thus, low temperatures and specific moisture conditions in high-latitude regions jointly result in relatively high soil DOC concentrations. However, substantial heterogeneity exists at regional and local scales. For instance, despite their similar latitudes, soil DOC concentrations in Northern Europe were significantly lower than in Siberia, primarily due to differences in climatic conditions. Northern Europe's maritime climate, with mild temperatures and evenly distributed precipitation, promotes higher microbial activity and accelerates organic matter decomposition. In contrast, Siberia's cold subarctic climate results in lower soil temperatures that limit microbial activity and slow organic matter decomposition, leading to greater DOC retention (Jin and Ma, 2021). Furthermore, soils in Siberia are often frozen, restricting oxygen supply and further inhibiting decomposition, thereby contributing to DOC accumulation (Raudina et al., 2022). Climatic conditions thus play a key role in explaining the significant differences in soil DOC concentrations between these regions.

Q29: Add recommendations for future sampling efforts

Re: It has been added to the part of 4.5 Limitations and predictive uncertainties.

Future research should enhance the collection of deep soil samples to address the current data scarcity

and more accurately quantify the DOC reserves across the entire soil profile. There is a particular need to increase sample collection in key regions such as Siberia and Africa.

Q30: L229-233: Update comparisons with more recent studies

Re: Done.

Q31: L235-237: Expand on ecosystem differences explanation

Re: It has been revised as “Tundra had the highest soil DOC concentration (Table 2). This can be attributed to low soil temperatures and limited microbial activity, which slow the decomposition of organic material and lead to higher soil DOC concentrations (Propster et al., 2023). In addition, prolonged soil freezing in tundra areas reduces evaporation and oxygen supply, further slowing organic decomposition. Soil DOC concentrations were also relatively high in grassland, forest, and shrub ecosystems because leaves, dead branches, and plant root exudates provide abundant organic C inputs (Cai et al., 2021). However, our results indicated that DOC concentrations in forest soils were consistently lower than in grasslands (Table 2). Grassland ecosystems often have higher plant diversity, including legumes and weeds, whose residue decomposition contributes to increased DOC concentrations (Perrot et al., 2023). In contrast, the cooler conditions in forest soils limit microbial activity and slow organic matter decomposition, reducing DOC consumption. Additionally, grassland soils tend to have better water conditions, promoting higher microbial activity and organic matter breakdown, thus increasing DOC concentrations (Deng et al., 2023). Differences in land use and management—forests being less disturbed while grasslands may be more frequently disturbed by grazing—can also influence soil organic matter decomposition and DOC levels. These combined factors of vegetation type, microbial activity, water conditions, and land use practices result in varying soil DOC concentrations between these two ecosystems. The lowest median soil DOC concentration appeared in cropland ecosystems, likely due to decreased soil organic matter inputs resulting from frequent tillage and harvesting, as well as accelerated DOC decomposition caused by tillage (Ren et al., 2024)”.

Q32: L250-254: Add mechanism explanations

Re: It has been revised as “In high-altitude regions, lower temperatures limit the metabolic activity of microorganisms, thereby slowing down the decomposition of soil DOC. Additionally, these areas typically receive more precipitation, which increases soil moisture and helps protect soil DOC from rapid decomposition. Soils with high clay content have a strong adsorption capacity, which can more

effectively retain soil DOC in the soil and reduce its loss. At the same time, clay provides a suitable habitat for microorganisms, affecting the structure and activity of microbial communities, and thus regulating the rate of soil DOC".

Q33: L278-299: Consider climate change implications

Re: Done. See Q28 for detailed responses.

Q34: Figure 1: add sample size for each ecosystem type

Re: Thanks for your good suggestion. Cropland is not included in Fig. 1c. Therefore, we have added the sample size of different ecosystems in the title.