https://doi.org/10.5194/essd-2024-343
Preprint. Discussion started: 29 August 2024

Open Access Earth System Science Data Discussions

Global patterns and drivers of soil dissolved organic carbon concentrations

Tianjing Ren and Andong Cai

Institute of Environment and Sustainable Development in Agriculture, Chinese Academy of Agricultural Sciences, Beijing, 100081, China

***Corresponding author:** caiandong@caas.cn

Tel: +86-10-82105615, Fax: +86-10-82105615

https://doi.org/10.5194/essd-2024-343
Preprint. Discussion started: 29 August 2024

Abstract

Dissolved organic carbon (DOC) is the most active carbon pool in soils, which plays critical roles in soil carbon cycling, plant productivity, and global climate change. An accurate assessment of the quantity of DOC in the soil is essential for the detailed elucidation of ecosystem functions and services. Nevertheless, the global driving factors and distribution of soil DOC remain inadequately quantified due to the scarcity of large-scale data. Here, a comprehensive global database of 12807 soil DOC concentrations derived from 975 target papers in the literature was compiled. Detailed geographic locations, climate, and soil properties were also recorded as predictors of soil DOC. Machine learning techniques were employed to assess the relative importance of various predictors in the determination of soil DOC concentrations, which were subsequently extended for their prediction on a global scale. The worldwide soil DOC concentration spanned a wide range (0.04 to 7859 mg kg-1), averaging 222.78 mg kg-1. The 12 selected variables (including soil properties, month, climate, and ecosystem) explained 65% of the variance in soil DOC concentrations. Elevation, soil clay, and soil organic carbon were three of the most important predictors. Global soil DOC concentration increased from the equator to the poles. The soil DOC stocks in the topsoil layer (0-30 cm) amounted to 12.17 Pg, with significant variations observed across different continents. These results are instrumental for informing strategies on soil management practices, ecosystem services, and the mitigation of climate change. Furthermore, our database can be combined with other carbon pools to explore the total soil carbon turnover and constrain Earth carbon models. The dataset is publicly available at https://doi.org/10.6084/m9.figshare.26379898 (Ren and Cai, 2024).

https://doi.org/10.5194/essd-2024-343
Preprint. Discussion started: 29 August 2024

1. Introduction

With global changes over the last few decades, terrestrial ecosystems, which serve as the fundamental safeguard for biodiversity and carbon sink on Earth, are becoming increasingly vital toward mitigating global climate warming (IPCC, 2014). Cumulatively, soil carbon pools constitute the largest carbon reservoirs of terrestrial ecosystems, which are three to four times greater than that of the ambient atmospheric carbon pool (Lal, 2004). Even minor fluctuations in soil carbon can have significant impacts on biogeochemical cycles and the global C balance. Dissolved organic carbon (DOC), which consists of simple organic acids and complex macromolecular substances, is recognized as the most active carbon pool in the soil. Currently, the portion of organic carbon that is water-soluble and can filter through 0.45 μm microporous filter membrane is referred to as DOC (Kalbitz et al., 2000; Zsolnay, 2003). Although soil DOC typically accounts for only < 2% of soil carbon pool, it provides a substantial source of carbon and energy for soil microorganisms, while playing a key role in soil carbon sequestration, transport, and stabilization mechanisms (Nakhavali et al., 2021; Ren et al., 2024). The lateral transport of DOC is crucial for linking terrestrial and aquatic ecosystems and plays a key role in the evaluation of terrestrial carbon budgets (Kindler et al., 2011; Sanderman & Amundson, 2008). Thus, an accurate assessment of soil DOC concentrations is vital due to its unique properties and roles, given its broad variations that can span up to three orders of magnitude (Nakhavali et al., 2020; Ren et al., 2024). Despite the significant variations in soil DOC concentrations, their global distribution has not yet been systematically quantified. Bridging this knowledge gap is essential for more accurately depicting the carbon cycle in Earth system models.

The soil DOC concentration depends on the dynamic balance between its sources (e.g., leachates from decomposing plant litter, plant root secretions, and microbial decomposition products) and losses (migration and microbial decomposition) (Bolan et al., 2011). Therefore, any factors that affect this dynamic balance would also influence the soil DOC concentrations. Extensive research has demonstrated that the soil DOC concentration is the outcome of climate, vegetation type, as well as soil properties (Chen et al., 2021; Guo et al., 2020; Smreczak & Ukalska-Jaruga, 2021). Each factor plays a distinct role in shaping soil DOC dynamics. For example, the climate, which is characterized by the annual mean temperature and precipitation, is typically recognized as a primary driving factor that influences the soil DOC concentrations (Kalbitz et al., 2000; Neff & Asner, 2001). Temperature and precipitation directly influence soil DOC concentrations by affecting microbial activities, organic matter

https://doi.org/10.5194/essd-2024-343
Preprint. Discussion started: 29 August 2024

Open Access Earth System Science Data Discussions

decomposition rates, its solubility and mobility, and indirectly modulate DOC dynamics by manipulating vegetation growth and soil structures (Andersson & Nilsson, 2001; Kalbitz et al., 2000). The type of vegetation impacts soil DOC concentrations mainly by affecting the input quantity and quality of organic matter (Guo et al., 2020). Together, climate and vegetation types have profound effects on soil biological, chemical, and physical properties, which are closely interconnected with the creation and decomposition of soil DOC (Camino‐Serrano et al., 2014). The relationships between soil DOC concentrations and environmental factors have been revealed based on local and regional scales. However, the relative importance of environmental factors that predict soil DOC concentrations on a global scale is still lacking, which impedes the development of effective strategies for the management of soil carbon and mitigation of climate change.

Accurate mapping of the soil DOC is essential for addressing pressing global challenges, including climate warming, food security, and eutrophication in aquatic systems (Guo et al., 2020; Langeveld et al., 2020). To the best of our knowledge, there are few global maps of the spatial distribution of soil DOC (Guo et al., 2020; Langeveld et al., 2020). However, these maps have subject to considerable uncertainties due to the limited data employed and the low interpretation rate. Firstly, there is a lack of valid observational data for Africa, South America, Eastern Europe, and Central Asia. Secondly, Guo (Guo et al., 2020) explained only 31% of the variations in the soil DOC using linear regression equations, while Langeveld (Langeveld et al., 2020) explained only 36%. In contrast to linear regression, machine learning has been extensively applied in research due to its capacities to automate feature extraction, handle large datasets, and recognize complex patterns, which offers significant advantages in terms of predictive accuracy and adaptive learning.

To address these challenges, we developed a comprehensive database of global soil DOC concentrations, comprising 12,807 samples from 975 published studies. Utilizing Random Forest algorithms, we quantified the relative importance of environmental factors, and further, predicted the soil DOC concentrations on a global scale. The special aims of this study were: (1) What are the global patterns of soil DOC concentrations? (2) What are the primary factors that control soil DOC concentrations on a global scale? (3) How large is total global soil DOC storage?

2. Material and method

2.1 Data sources and processing

https://doi.org/10.5194/essd-2024-343
Preprint. Discussion started: 29 August 2024

Open Access Earth System Science Data Discussions

Publication search for this study was performed using Google Scholar (https://scholar.google.com), the Web of Science (http://apps.webofknowledge.com), and the China Knowledge Resource Integrated Database (http://www.cnki.net/) using the following search terms: (dissolved organic carbon OR dissolved organic matter OR "DOC" OR "DOM") AND soil, up to December 2022. The specific data flow through the different phases for the selected papers is shown in Fig. S1. To ensure a standardized and bias-minimized dataset, the following inclusion criteria were applied: (1) Data must be from terrestrial ecosystems, excluding oceans and rivers; (2) Only the topsoil layer data (0-30 cm) were used; (3) Duplicate results from different articles were recorded only once; (4) Soils included agricultural soils that were affected by human activities through tilling and fertilization etc., but did not cover industrial and urban soils. Data presented solely in figures were extracted using the digitizer function of Origin 2019 software.

Based on these criteria, a total of 12807 observations of soil DOC were compiled from 975 publications. Additional data included specifics of the experimental sites (longitude, latitude, and altitude), climatic conditions (mean annual temperature (MAT) and mean annual precipitation (MAP)), biomes (e.g., wetland，forest，shrubland，tundra，grassland，and cropland) and soil physical and chemical properties (e.g., soil organic carbon, texture, and pH) (Table 1). These environmental factors are used as predictors. When those environmental factors were missing within the original publication, the missing data were extracted from the grid dataset according to geographic coordinates of observed site (Table S1). This study sites spanned a wide range of latitudes (-64.81° to 78.85°) and longitudes (-159.66° to 175.95°) (Table 1). This database encompassed a large gradient of climate regimes, with MAT from −11.16 to 28.00°C and MAP from 30 to 4200 mm.

2.2 Data standardization

In our database, the DOC concentrations were quantified using a mix of physical and chemical techniques. Physical methods included soil solution collection using lysimeters or ceramic suction. Chemical methods employed various solvents like distilled water, potassium chloride (KCl), or potassium sulfate (K_2SO_4) as described by Li et al. (2018). Over 74.32% of the DOC was determined using chemical techniques, which highlighted their reliability. For consistency, the DOC values derived from physical approaches was converted to chemical method values using the following equation:

$$DOC_{soil} = (DOC_{solution} \times V \times 1000)/W \times [1/(V \times (1-W) \times BD \times 1000000)] \quad (1)$$

https://doi.org/10.5194/essd-2024-343
Preprint. Discussion started: 29 August 2024

where, DOC_{soil} represents soil DOC concentration determined by chemical methods (mg g^{-1}); $DOC_{solution}$ is the concentration measured by physical methods (mg L^{-1}); W denotes the volumetric soil moisture (m^3 m^{-3}); V is the volume of the soil column for solution extraction (m^3); and BD is the soil bulk density (g cm^{-3}). The factor 1000 converts m^3 to L, and 1000000 converts m^3 to cm^3 following the protocol established by Guo (Guo et al., 2020). This standardization allowed for a consistent comparison and analysis of the DOC data across various studies.

2.3 Predictive modeling

The driving factors of soil DOC concentrations were divided into four categories (climate, ecosystem, soil properties, and observation time). The soil properties included physical (clay, sand, bulk density, and depth), chemical (SOC, pH), biological (microbial biomass carbon) attributes. The observation time was represented by month. Climate referred to MAT, MAP, and elevation. Ecosystems encompassed wetland, forest, shrubland, tundra, grassland, and cropland. In predictive models, correlated predictors may substitute for each other, such that their importance will be shared, which results in an estimated importance that is less than the true value. Consequently, the soil total nitrogen, silt, and aridity index were not included as they were correlated with the soil organic carbon, sand, and MAP, respectively (Fig. S2). Further, some variables were not included due to rarely report in target paper.

To develop and optimize a predictive model for soil DOC an array of regression methods was employed, which encompassed three linear and four nonlinear approaches (Table S2). The linear regression methods included a least absolute shrinkage and selection operator (LEAPS), elastic net (ENET), and standard linear modeling (LM) to identify the most important predictor variable in a regression model, while minimizing the risk of overfitting. The nonlinear regression methods included the random forest (RF) algorithm, boosted tree (BOOSTED), bagged tree (Bagged), and cubist (CUBIST) models. Each model was equipped with intrinsic feature selection processes and was fine-tuned to improve accuracy and control complexity. During the optimization phase, various actions were implemented; LEAPS models were educated to accommodate the highest count of variables. To discipline the models, the penalty for feature condensation (diminishing the role of less impactful variables in the resultant linear formula) varied between 0 and 0.1, incremented by 0.01. RF models' growth was capped at a maximum of 1,000 trees, and the model's predictors were restricted to a third of the possible maximum, ensuring a balance between complexity and manageability. BOOSTED models underwent training with a tree count ranging from ten to a hundred, where each tree had a node range of one to seven. They incorporated a shrinkage rate of either 0.01 or 0.1, and a maximum size limit set to five, optimizing the models' learning process. CUBIST model utilized a sequence of

https://doi.org/10.5194/essd-2024-343
Preprint. Discussion started: 29 August 2024

neighboring values from 1 to 9 with increments of 2, alongside community sizes spanning 1 to 100, to refine its predictive accuracy. In every instance, the models were evaluated using Monte Carlo cross-validation, which involved 100 iterations of data resampling with an 80/20 split between training and validation datasets, ensuring an accurate estimation of model uncertainty and safeguarding against over-fitting. 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 (Table S2). The relative RMSE, a measure of the estimation uncertainty for soil DOC, was determined by dividing the error's magnitude by the overall average soil DOC value. The nonlinear models (R^2 = 0.42-0.65; root mean square error (RMSE) = 250-332) outperformed the linear models (R^2 = 0.101-0.108; RMSE = 410-427) (Table S2). The RF model distinguished itself with the lowest RMSE within a standard deviation range, and the model was then selected for subsequent analyses focusing on variable importance (Fig. S3). Consequently, the relative importance of driving the soil DOC and the global map of soil DOC were the averaged values of the RF model results.

To evaluate the impacts of independent variables on the soil DOC, a variable importance analysis was conducted using permutation variable importance measurements (Fig. 2). This analysis was performed utilizing the variable importance tool integrated into the R packages for the RF model that exhibited the highest accuracy and predictive quality. In essence, this method assessed prediction errors within the model by calculating mean square errors for each regression tree. The models' variable importance scores assessed the influence of predictor variables on the outcomes. For enhanced comparability of all model inputs, the independent environmental variables were scaled to a 0 to 100% range, reflecting their proportional contribution to the model's predictions.

Partial dependence analyses were employed to test the relationships between the predicted soil DOC and independent variables across the entire spectrum of potential values considered in the RF model (Fig. 3). In essence, this approach provided insights into the global relationships between the independent variables and predicted outcomes. The focus was set solely on the effects of the targeted independent variables by eliminating the influences of other independent variables. Partial dependence analyses, along with their graphical representations known as partial dependence plots, provided insight into the average marginal effect of one or more independent variables on a machine learning model's predictions within a defined value scope, offering a more nuanced view than assessing the overall relative importance of an independent variable. For instance, partial dependence plots can expose whether the connection between a predicted variable and an independent control is linear, monotonic, or complex. The

https://doi.org/10.5194/essd-2024-343
Preprint. Discussion started: 29 August 2024

curvature and inflection points of the partial dependence plot curve help us to decipher and pinpoint areas where an independent variable exerts a notably strong and immediate effect on the forecasted outcome. Additionally, it can indicate where the variable's influence is more subtle, potentially mediated through its effects on other independent variables. To facilitate the interpretation of the partial dependence plots, the x-axis for the standardized value was reported, which ensured a clear progression from low to high values in all curves.

2.4 Global soil DOC mapping

The global distribution of the soil DOC and the relative uncertainties of predictions were generated (Figs. 4, S5). These maps were derived by utilizing our DOC dataset in conjunction with the RF model, which incorporated the global climate, vegetation, and soil-rasterized datasets (Table S1). We generated factor maps from the key input variables, focusing on the 12 distinct variables associated with each raster cell. Subsequently, the factor maps were employed to derive a spatially detailed global map of soil DOC. For global scale mapping, the driving factors were initially processed at a 0.05° resolution to calculate the soil DOC values. Areas that did not meet the following criteria were excluded from our prediction: (1) absence of data for any essential predictors, (2) soil order and biomes not aligning with the previously discussed aggregated land use systems, or (3) locations in climate zones outside the scope of our model's focus. To evaluate the uncertainty associated with map creation due to data resampling and any unexplained variability unaccounted for by the independent variables, we analyzed finer resolution (5 km^2) grids in regions where driving factors were accessible at this detailed level. This analysis illuminated the overall uncertainty inherent in our global soil DOC estimation. A map representing the relative prediction uncertainty was crafted, showcasing the standard deviation in relation to the mean of the predictions. The standard deviation, indicative of the dispersion in potential predictions, was derived from the decision tree model's structure after 500 iterations of the model.

3. Results

3.1 Soil DOC concentrations in different ecosystems globally

A total of 12,807 soil DOC observations were compiled from 975 publications, which spanned six continents as well as major biomes and terrestrial ecosystems (Fig. 1), and the database conformed to a normal distribution (Fig. 1b). The global soil DOC concentrations varied between 0.04 and 7859 mg kg^{-1}. The global average, median, and standard deviation were 222.78, 101.01, and 445.78 mg kg^{-1}, respectively (Table 2). The concentrations of soil DOC

https://doi.org/10.5194/essd-2024-343
Preprint. Discussion started: 29 August 2024

varied across different ecosystems. Tundra had the highest soil DOC concentration (470.78 mg kg^{-1}), while shrubland had the lowest (160.24 mg kg^{-1}). The average soil DOC concentrations for grassland, forest, wetland, and cropland were 327.77, 256.18, 218.53, and 165.98 mg kg^{-1}, respectively (Table 2).

3.2 Model performance and drivers of soil DOC concentrations

Random forest model accounted for 65% of the variability in soil DOC concentrations across all sites, with the lowest RMSE compared with other models (Fig. 2, Table S2). The most important categories of predictors for soil DOC concentrations were climate and soil properties, with elevation and the soil clay content emerging as the most significant. Although less influential, other predictors were nonetheless considered, with soil organic carbon and soil pH having the most notable effects (Fig. 2a). Although the mean annual precipitation and temperature, microbial biomass carbon, bulk density, sand, depth, month, and ecosystem affected soil DOC concentrations, their relative contributions were lower than aforementioned four predictors (Fig. 2). Partial dependence analysis showed similar results to Pearson correlation analysis (Fig. S4) and indicated that there was a positive correlation between the soil DOC and both the elevation and soil organic carbon (Fig. 3g). Conversely, the soil DOC was negatively correlated with mean annual temperature and soil pH (Fig. 3h).

3.3 Global soil DOC patterns

Our predicted global soil DOC mapping implied that there was a significant spatial heterogeneity of soil DOC concentrations (Fig. 4a). This revealed a latitudinal pattern that soil DOC concentrations increased from the equator to poles (Fig. 4b). High soil DOC concentrations were found in high-altitude plateaus and mountain ranges at low latitude (e.g., Andes, African Highlands, West Indies) (Fig. 4a). The global average soil DOC concentration was 237.56 mg kg^{-1} (Table 3), while the soil DOC stock in the topsoil (0-30 cm) was 12.17 Pg.

Asia had the highest soil DOC concentration (274.43 mg kg^{-1}) followed by North America (263.63 mg kg^{-1}). Next were Europe and South America (227.34 and 215.81 mg kg^{-1}, respectively), with Oceania and Africa having the lowest soil DOC concentrations (198.13 and 186.35 mg kg^{-1}, respectively). For predicted soil DOC stocks, Asia and North America remained in first and second place (4.8 and 2.45 Pg, respectively). Despite the marginal predicted soil DOC concentrations in Africa, its predicted soil DOC stocks ranked third (2.07 Pg) due to its vast area. South

https://doi.org/10.5194/essd-2024-343
Preprint. Discussion started: 29 August 2024

America was in fourth place with a predicted soil DOC stock of 1.37 Pg. Finally, Europe and Oceania showed the lowest predicted soil DOC stocks (0.88 and 0.59 Pg, respectively).

4 Discussions

4.1 Variations in soil DOC between ecosystems

Given the substantial number of measurements included in our study (12,807 observations), the range of soil DOC concentrations (0.04-7859 mg kg^{-1}) was broader than that reported by Guo (3,869 observations) (Guo et al., 2020). Our reported global average soil DOC concentration was 222.78 mg kg^{-1} (Table 2), in contast to Guo's reported average of only 77.39 mg kg^{-1}. For different ecosystems, the soil DOC concentrations of wetlands, tundra, and shrublands in our study aligned with those of previous research (Guo et al., 2020), which was primarily due to the relatively lower number of observations for these ecosystems in comparison withothers, with tundra comprising only 1% of our database (Guo et al., 2020). However, significant differences were found in forests, grasslands, and croplands compared with Guo's data. For instance, our average soil DOC concentration for croplands was 165.98 mg kg^{-1}, while Guo reported only 60.58 mg kg^{-1}. This discrepancy was due to Guo's database including only 13% cropland observations, whereas our cropland observations are approximately ten times larger (Guo et al., 2020). However, our results consistently indicated that DOC concentrations in forest soils were lower than in grasslands, with tundra showing the highest DOC levels (Table 2) (Guo et al., 2020). This was due to the higher lignin content in forests, which reduces the quality of plant litter, hinders microbial decomposition, and releases less DOC (Wang et al., 2015). For tundra, besides low microbial activities in permafrost due to low temperatures, anaerobic conditions from soil oversaturation severely limit microbial activities and growth, reduce decomposition rates, and increase the DOC (Boddy et al., 2008; Petrone, 2005). Despite the frequent addition of nutrients in croplands, the DOC concentrations remained lower than expected. Intensive anthropogenic activities, such as management practices and frequent harvesting induced the significant loss of soil organic matter, which translated to reduced DOC (Guo et al., 2020; Li et al., 2019; Ren et al., 2024). In summary, our study built on preceding work by incorporating a more extensive dataset that better represented the heterogeneous conditions found globally.

4.2 Effects of climate and controlled soil properties on soil DOC concentrations

https://doi.org/10.5194/essd-2024-343
Preprint. Discussion started: 29 August 2024

Open Access Earth System Science Data Discussions

The two most critical predictors of soil DOC concentrations were climate and soil properties, with elevation and soil clay content being the two most significant factors (Fig. 3). As the elevation gradient increase, temperatures generally decrease, which can constrain microbial metabolic rates and reduce the decomposition of organic matter, which leads to additional organic carbon being retained in the soil as DOC (Li et al., 2023; Nottingham et al., 2019; Wei et al., 2024). Typically, high-altitude regions host specific vegetation types with longer growth cycles and more litterfall (Pesántez et al., 2018; Wei et al., 2024). These plant residues decompose to SOC, a portion of which converts to DOC. Consequently, differences in the vegetation type and productivity also influence the soil DOC concentrations (Camino‐Serrano et al., 2014; Rahbek et al., 2019). We also found that forest and grassland sites above 2000 m (which constituted 73% of the high DOC observations) were significant contributors. High-altitude regions often experience distinct precipitation patterns and soil moisture conditions compared with lower elevations (Li et al., 2023). Higher precipitation and lower evaporation rates may result in the greater dissolution and leaching of organic matter, thereby increasing DOC concentrations in the soil (He et al., 2021; Lu et al., 2019). High-altitude areas are generally less frequented by humans, which may assist in the preservation of the DOC in the soil through the prevention of disturbances and losses. Our results also indicated that soils in low-latitude plateaus and mountain ranges (e.g., Tibetan Plateau, Andes, African Highlands, and West Indies) exhibited higher DOC concentrations (Fig. 4a). The impacts of the soil clay content on DOC concentrations are complex, which occurred primarily through adsorption, water retention, microbial activities, and organic matter protection mechanisms (Kaiser & Zech, 2000; Singh et al., 2017). Generally, a high clay content tends to stimulate the accumulation of soil DOC through the adsorption and stabilization of organic matter (Gmach et al., 2019; Kalbitz et al., 2000). Furthermore, the effects of SOC and soil pH on DOC should not be overlooked (Fig. 2a). SOC serves as the main source of DOC, where higher SOC generally implies that more DOC can be released into the soil through microbial metabolism (Kalbitz et al., 2000; Neff & Asner, 2001). Variations in the soil pH can affect the charge of soil colloids, thereby altering their adsorption-desorption mechanisms for DOC, which affects its solubility in the soil (Andersson & Nilsson, 2001; Cheng et al., 2020; Kaiser et al., 2005). In summary, the soil DOC concentration is the result of interactions between the soil and climate, biological, chemical, physical processes, and human influences at various spatial and temporal scales, with each factor playing a unique role in shaping DOC dynamics.

4.3 Global patterns of soil DOC

https://doi.org/10.5194/essd-2024-343
Preprint. Discussion started: 29 August 2024

Using our soil DOC concentration dataset, we quantified the soil DOC concentrations (0-30 cm) in terrestrial ecosystems, identified their key driving factors, and made global predictions. Global DOC stocks in the topsoil are estimated at 12.17 Pg C, accounting for 0.775% of the global soil organic carbon, which is significantly higher than previous estimates (Guo et al., 2020). Our predictions indicated that soil DOC concentrations decreased significantly with lower latitudes, particularly in the Northern Hemisphere. Previous global maps of soil DOC concentrations failed to capture this latitudinal trend, which was likely due to their limited spatial coverage (Guo et al., 2020; Langeveld et al., 2020). Our predicted map shows that the soil DOC concentrations increased with latitude. This trend was attributed to lower temperatures, specific vegetation types, higher soil moisture, and reduced human activities at higher latitudes (Camino‐Serrano et al., 2014; Lapierre et al., 2015). However, there was substantial heterogeneity at regional and local scales. For instance, despite being at similar latitudes, soil DOC concentrations in Northern Europe were significantly lower than in Siberia, which we surmised was primarily due to differences between the maritime climate of Northern Europe and the cold subarctic climate of Siberia. Regional variations in soil DOC concentrations might be related to topographic condition. Higher soil DOC concentrations on the Tibetan Plateau compared to Eastern China might result from the high elevation and low MAT in the plateau (Fig. 4a). In contrast, lower DOC levels in Arctic regions was reported, which might have been due to their omission of DOC concentration in the soil and dry or frozen soil (Langeveld et al., 2020). The predictive model offered higher accuracy in estimating the global soil DOC storage (Fig. 3). This advantage stemmed from our comprehensive dataset, which included DOC concentrations in both dry soil and soil solutions, which provided a robust data foundation for global soil DOC predictions. Additionally, we employed the optimal model for predicting the global soil DOC by comparing various linear and non-linear models.

4.4 Limitations and predictive uncertainties

Although we compiled a comprehensive global soil DOC concentration dataset, identified key drivers, and made a global prediction, our study had certain limitations. First, certain ecosystems remained underrepresented; for instance, tundra accounted for only 1% of our database, while shrublands, grasslands, and wetlands collectively constituted only 21%. This underrepresentation may reduce the accuracy of predictions for different ecosystems. Second, although we considered the subsoil at the beginning of dataset, we did not explore this further due to the limited availability of data and considerations of predictive accuracy. We intend to continue expanding the subsoil

https://doi.org/10.5194/essd-2024-343
Preprint. Discussion started: 29 August 2024

DOC database in future work. Third, 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. Fourth, despite employing advanced machine learning methods with multiple predictors to predict the global soil DOC, 35% of soil DOC concentration variability remains unexplained. However, these limitations also highlighted areas for future soil DOC research.

5 Data availability

The global soil DOC in this study and raw dataset of driving factors can be downloaded at https://doi.org/10.6084/m9.figshare.26379898 (Ren and Cai, 2024).

6 Conclusions

Through the development of a comprehensive soil DOC dataset, we quantified soil DOC concentrations in terrestrial ecosystems, identified their driving factors, and made global predictions. Subsequent to comparing multiple predictive models, we selected the Random Forest model as the best performer for mapping soil DOC concentrations. The results indicated that tundra exhibited the highest DOC concentrations, while shrubland and cropland soils had relatively lower concentrations. Climate factors (elevation) and soil properties (clay content, SOC, pH) jointly regulated the DOC variations. The predicted that the soil DOC concentration increased significantly from the equator to the poles, and estimated the DOC stocks in the topsoil of terrestrial ecosystems was 12.17 Pg. The global soil DOC database we created will serve as a critical resource for future research, while enhancing our understanding of the roles of soil in the global carbon cycle. This database provides valuable data support for climate change research, ecosystem management, agricultural sustainability, environmental policymaking, and the improvement of biogeochemical models. This will aid in addressing soil degradation, improving food security, and tackling global environmental challenges.

Author contributions

Andong Cai designed this study. Tianjing Ren collected the data. Tianjing Ren and Andong Cai discussed analyzing methods. Andong Cai conducted the analysis. Tianjing Ren drafted the manuscript. All authors discussed the results and contributed to the manuscript.

Competing interests

The contact author has declared that neither they have any competing interests.

Acknowledgements

We would like to thank Frank Boehm at NanoApps Consulting2341York Ave. Vancouver, BC, Canada for his assistance with English language and grammatical editing.

Financial support

This work was financially supported by the National Key Research and Development Program of China (2022YFD2300500).

References

Andersson, S., & Nilsson, S. I. (2001). Influence of pH and temperature on microbial activity, substrate availability of soil-solution bacteria and leaching of dissolved organic carbon in a mor humus. Soil Biology and Biochemistry, 33(9), 1181-1191. https://doi.org/10.1016/S0038-0717(01)00022-0

Boddy, E., Roberts, P., Hill, P. W., Farrar, J., & Jones, D. L. (2008). Turnover of low molecular weight dissolved organic C (DOC) and microbial C exhibit different temperature sensitivities in Arctic tundra soils. Soil Biology and Biochemistry, 40(7), 1557-1566. https://doi.org/10.1016/j.soilbio.2008.01.030

Bolan, N. S., Adriano, D. C., Kunhikrishnan, A., James, T., McDowell, R., & Senesi, N. (2011). Dissolved organic matter: biogeochemistry, dynamics, and environmental significance in soils. Advances in agronomy, 110, 1-75. https://doi.org/10.1016/B978-0-12-385531-2.00001-3

Camino-Serrano, M., Gielen, B., Luyssaert, S., Ciais, P., Vicca, S., Guenet, B., Vos, B. D., Cools, N., Ahrens, B., Altaf Arain, M., Borken, W., Clarke, N., Clarkson, B., Cummins, T., Don, A., Pannatier, E. G., Laudon, H., Moore, T., Nieminen, T. M., . . . Janssens, I. (2014). Linking variability in soil solution dissolved organic carbon to climate, soil type, and vegetation type. Global Biogeochemical Cycles, 28(5), 497-509. https://doi.org/10.1002/2013gb004726

https://doi.org/10.5194/essd-2024-343
Preprint. Discussion started: 29 August 2024

Chen, H., Kong, W., Shi, Q., Wang, F., He, C., Wu, J., Lin, Q., Zhang, X., Zhu, Y. G., Liang, C., & Luo, Y. (2021). Patterns and drivers of the degradability of dissolved organic matter in dryland soils on the Tibetan Plateau. Journal of Applied Ecology. https://doi.org/10.1111/1365-2664.14105

Cheng, X., Hou, H., Li, R., Zheng, C., & Liu, H. (2020). Adsorption behavior of tetracycline on the soil and molecular insight into the effect of dissolved organic matter on the adsorption. Journal of Soils and Sediments, 20(4), 1846-1857. https://doi.org/10.1007/s11368-019-02553-7

Gmach, M. R., Cherubin, M. R., Kaiser, K., & Cerri, C. E. P. (2019). Processes that influence dissolved organic matter in the soil: a review. Scientia Agricola, 77. https://doi.org/10.1590/1678-992X-2018-0164

Guo, Z., Wang, Y., Wan, Z., Zuo, Y., He, L., Li, D., Yuan, F., Wang, N., Liu, J., & Song, Y. (2020). Soil dissolved organic carbon in terrestrial ecosystems: Global budget, spatial distribution and controls. Global Ecology and Biogeography, 29(12), 2159-2175. https://doi.org/10.1111/geb.13186

He, X., Augusto, L., Goll, D. S., Ringeval, B., Wang, Y., Helfenstein, J., Huang, Y., Yu, K., Wang, Z., Yang, Y., & Hou, E. (2021). Global patterns and drivers of soil total phosphorus concentration. Earth System Science Data, 13(12), 5831-5846. https://doi.org/10.5194/essd-13-5831-2021

IPCC. (2014). Climate Change 2014: Synthesis Report. In Core Writing Team, R. K. Pachauri & L. A. Meyer (Eds.), Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Geneva, Switzerland: IPCC. (2). https://epic.awi.de/id/eprint/37530/

Kaiser, K., Guggenberger, G., Haumaier, L., & Zech, W. (2005). Dissolved organic matter sorption on sub soils and minerals studied by 13C-NMR and DRIFT spectroscopy. European Journal of Soil Science, 48(2), 301-310. https://doi.org/10.1111/j.1365-2389.1997.tb00550.x

Kaiser, K., & Zech, W. (2000). Dissolved organic matter sorption by mineral constituents of subsoil clay fractions. Journal of Plant Nutrition and Soil Science, 163(5), 531-535. https://doi.org/10.1002/1522-2624(200010)163:5%3C531::AID-JPLN531%3E3.0.CO;2-N

Kalbitz, K., Solinger, S., Park, J.-H., Michalzik, B., & Matzner, E. (2000). Controls on the dynamics of dissolved organic matter in soils: a review. Soil Science, 165(4), 277-304. https://doi.org/10.1097/00010694-200004000-00001

https://doi.org/10.5194/essd-2024-343
Preprint. Discussion started: 29 August 2024

Kindler, R., Siemens, J., Kaiser, K., Walmsley, D. C., Bernhofer, C., Buchmann, N., Cellier, P., Eugster, W., Gleixner, G., & Grũnwald, T. (2011). Dissolved carbon leaching from soil is a crucial component of the net ecosystem carbon balance. Global Change Biology, 17(2), 1167-1185. https://doi.org/10.1111/j.1365-2486.2010.02282.x

Lal, R. (2004). Soil carbon sequestration impacts on global climate change and food security. science, 304(5677), 1623-1627. https://doi.org/10.1126/science.1097396

Langeveld, J., Bouwman, A. F., van Hoek, W. J., Vilmin, L., Beusen, A. H. W., Mogollón, J. M., & Middelburg, J. J. (2020). Estimating dissolved carbon concentrations in global soils: a global database and model. Sn Applied Sciences, 2(10). https://doi.org/10.1007/s42452-020-03290-0

Lapierre, J. F., Seekell, D. A., & Del Giorgio, P. A. (2015). Climate and landscape influence on indicators of lake carbon cycling through spatial patterns in dissolved organic carbon. Global Change Biology, 21(12), 4425-4435. https://doi.org/10.1111/gcb.13031

Li, mengfan, Wang, J., Guo, D., Yang, R., & Fu, H. (2019). Effect of land management practices on the concentration of dissolved organic matter in soil: A meta-analysis. Geoderma, 344, 74-81. https://doi.org/10.1016/j.geoderma.2019.03.004

Li, J., Wu, B., Zhang, D., & Cheng, X. (2023). Elevational variation in soil phosphorus pools and controlling factors in alpine areas of Southwest China. Geoderma, 431. https://doi.org/10.1016/j.geoderma.2023.116361 https://doi.org/10.1016/j.geoderma.2023.116361

Li, S., Zheng, X., Liu, C., Yao, Z., Zhang, W., & Han, S. (2018). Influences of observation method, season, soil depth, land use and management practice on soil dissolvable organic carbon concentrations: A meta-analysis. Science of the Total Environment, 631-632, 105-114. https://doi.org/10.1016/j.scitotenv.2018.02.238

Lu, S., Xu, Y., Fu, X., Xiao, H., Ding, W., & Zhang, Y. (2019). Patterns and drivers of soil respiration and vegetation at different altitudes in Southern China. Applied Ecology & Environmental Research, 17(2). https://doi.org/10.15666/aeer/1702_30973106

Nakhavali, M., Lauerwald, R., Regnier, P., Guenet, B., Chadburn, S., & Friedlingstein, P. (2020). Leaching of dissolved organic carbon from mineral soils plays a significant role in the terrestrial carbon balance. Glob Chang Biol. https://doi.org/10.1111/gcb.15460

https://doi.org/10.5194/essd-2024-343
Preprint. Discussion started: 29 August 2024

Nakhavali, M., Lauerwald, R., Regnier, P., Guenet, B., Chadburn, S., & Friedlingstein, P. (2021). Leaching of dissolved organic carbon from mineral soils plays a significant role in the terrestrial carbon balance. Global Change Biology, 27(5), 1083-1096. https://doi.org/10.1111/gcb.15460

Neff, J. C., & Asner, G. P. (2001). Dissolved organic carbon in terrestrial ecosystems: synthesis and a model. Ecosystems, 4(1), 29-48. https://doi.org/10.1007/s100210000058

Nottingham, A. T., Baath, E., Reischke, S., Salinas, N., & Meir, P. (2019). Adaptation of soil microbial growth to temperature: Using a tropical elevation gradient to predict future changes. Glob Chang Biol, 25(3), 827-838. https://doi.org/10.1111/gcb.14502

Pesántez, J., Mosquera, G. M., Crespo, P., Breuer, L., & Windhorst, D. (2018). Effect of land cover and hydro-meteorological controls on soil water DOC concentrations in a high-elevation tropical environment. Hydrological Processes, 32(17), 2624-2635. https://doi.org/10.1002/hyp.13224

Petrone, K. C. (2005). Export of carbon, nitrogen and major solutes from a boreal forest watershed: The influence of fire and permafrost. University of Alaska Fairbanks.

Rahbek, C., Borregaard, M. K., Colwell, R. K., Dalsgaard, B., Holt, B. G., Morueta-Holme, N., Nogues-Bravo, D., Whittaker, R. J., & Fjeldså, J. (2019). Humboldt's enigma: What causes global patterns of mountain biodiversity? Science, 365(6458), 1108-1113. https://www.science.org/doi/10.1126/science.aax0149

Ren, T., Ukalska-Jaruga, A., Smreczak, B., & Cai, A. (2024). Dissolved organic carbon in cropland soils: A global meta-analysis of management effects. Agriculture, Ecosystems & Environment, 371, 109080. https://doi.org/10.1016/j.agee.2024.109080

Ren, T., & Cai, A. (2024). Global patterns and drivers of soil dissolved organic carbon concentrations. figshare [data set]. https://doi.org/https://doi.org/10.6084/m9.figshare.26379898

Sanderman, J., & Amundson, R. (2008). A comparative study of dissolved organic carbon transport and stabilization in California forest and grassland soils. Biogeochemistry, 89, 309-327. https://doi.org/10.1007/s10533-008-9221-8

Singh, M., Sarkar, B., Hussain, S., Ok, Y. S., Bolan, N. S., & Churchman, G. J. (2017). Influence of physico-chemical properties of soil clay fractions on the retention of dissolved organic carbon. Environmental Geochemistry and Health, 39, 1335-1350. https://doi.org/10.1007/s10653-017-9939-0

https://doi.org/10.5194/essd-2024-343
Preprint. Discussion started: 29 August 2024

Smreczak, B., & Ukalska-Jaruga, A. (2021). Dissolved organic matter in agricultural soils. Soil Science Annual. https://doi.org/10.37501/soilsa/132234

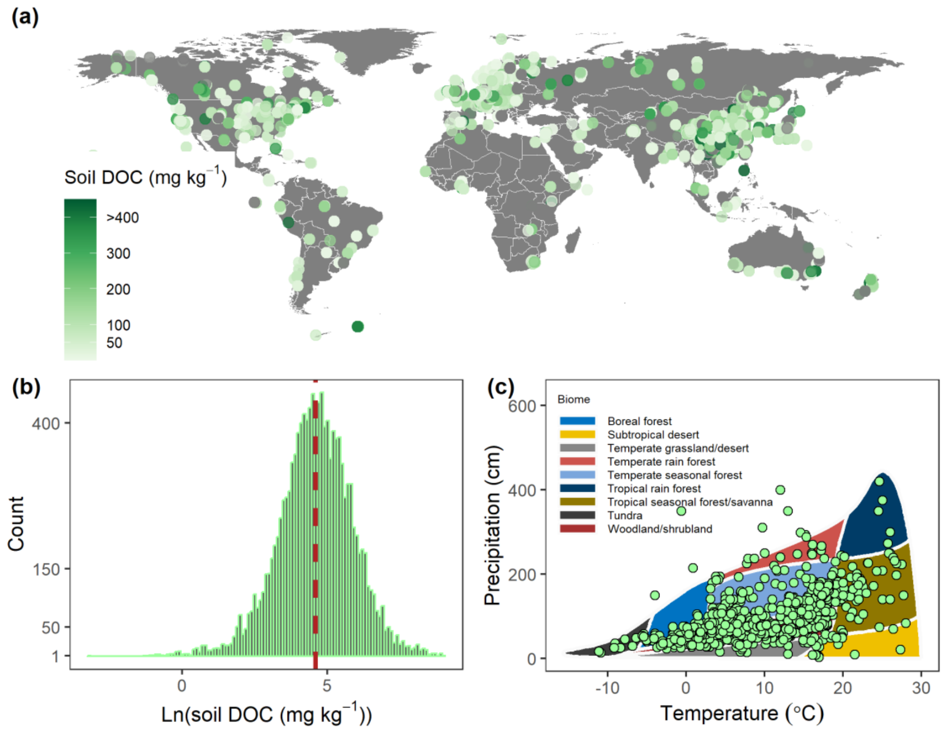
Wang, Y.-L., Chang-Ming, Y., Li-Min, Z., & Heng-Zhao, C. (2015). Spatial distribution and fluorescence properties of soil dissolved organic carbon across a riparian buffer wetland in Chongming Island, China. Pedosphere, 25(2), 220-229. https://doi.org/10.1016/S1002-0160(15)60007-8

Wei, D., Tao, J., Wang, Z., Zhao, H., Zhao, W., & Wang, X. (2024). Elevation-dependent pattern of net CO(2) uptake across China. Nature Communications, 15(1), 2489. https://doi.org/10.1038/s41467-024-46930-4

Zsolnay, Á. (2003). Dissolved organic matter: artefacts, definitions, and functions. Geoderma, 113(3-4), 187-209. https://doi.org/10.1016/s0016-7061(02)00361-0

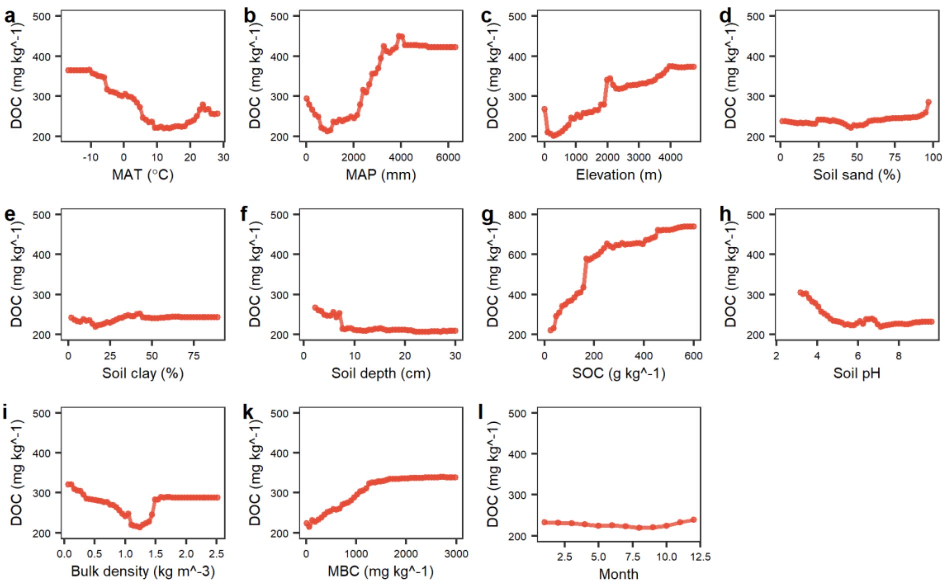
https://doi.org/10.5194/essd-2024-343
Preprint. Discussion started: 29 August 2024

Figure 1 Global distribution of soil dissolved organic carbon (DOC) concentration according to our site-level dataset. The dataset contains 12807 sets of data (**a, b**), which covers major terrestrial biomes (**c**). The dashed red line within the subplot (**b**) signifies the average soil DOC concentration, which is 223 mg kg^{-1}.



https://doi.org/10.5194/essd-2024-343
Preprint. Discussion started: 29 August 2024

Figure 2 Partial dependence of predictors from random forest algorithm. Soil dissolved organic carbon (DOC) concentration in relation to mean annual temperature (MAT), mean annual precipitation (MAP), elevation, soil sand content, soil clay content, soil depth, soil organic carbon (SOC) content, soil pH, bulk density, microbial biomass carbon content (MBC), and month (**a, b, c, d, e, f, g, h, i, k, l,** respectively).



https://doi.org/10.5194/essd-2024-343
Preprint. Discussion started: 29 August 2024

Figure 3 Result of the random forest model predicting soil dissolved organic carbon (DOC) concentration. (**a**) The relative importance of predictors in the random forest model. (**b**) Predicted vs. observed soil DOC concentration. The dashed line indicates the 1:1 line and the blue line indicates the regression line between predicted and observed values.

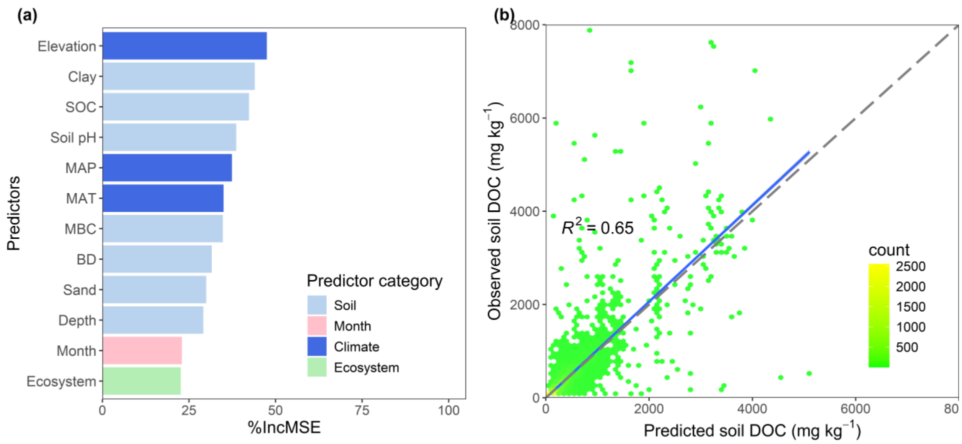
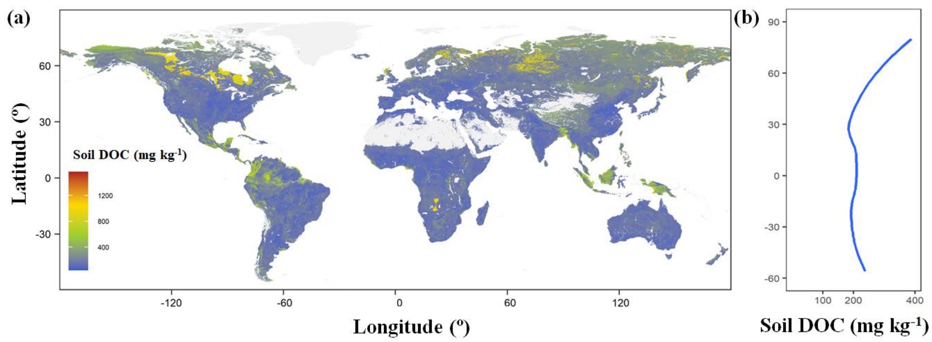


Figure 4 Prediction of soil dissolved organic carbon (DOC) concentration in global ecosystems. (**a**) Global map of predicted soil DOC concentration. (**b**) Latitudinal patterns of soil DOC concentration. Blue line indicates the locally weighted regressions between latitude and soil DOC concentration in the predicted global map. Values in the predicted map reflect soil DOC concentration within a grid cell resolution of 0.05° × 0.05°. A value in the grid is the averaged from the result of random forest model.



https://doi.org/10.5194/essd-2024-343
Preprint. Discussion started: 29 August 2024

Open Access Earth System Science Data Discussions

Table 1. Variables information of soil dissolved organic carbon dataset in global terrestrial ecosystems. n/a refers to values that are not applicable.

Variables	Description	Unit	Number	Range	Mean
No.	Unique identification number of each record	n/a	12807	1 to 12807	6404
Latitude	Latitude of study site	°	12807	-64.81 to 78.85	34.89
Longitude	Latitude of study site	°	12807	-159.66 to 175.95	107.05
MAT	Mean annual temperature	℃	9948	-11.16 to 28.00	11.84
MAP	Mean annual precipitation	mm	10325	30 to 4200	1071
Elevation	Altitude of study site	m	5578	4 to 4730	881
Ecosystems	Community by the dominant plant species		7	n/a	n/a
Soil sand	Soil sand content	%	4062	1 to 98	45
Soil silt	Soil silt content	%	4025	1 to 95	33
Soil clay	Soil clay content	%	4316	0 to 89	22
Soil depth	Mean depth of soil sample	cm	12807	0.53 to 30.00	11.36
SOC	Soil organic carbon	g kg^{-1}	9136	0.23 to 598.50	38.74
TN	Soil total nitrogen	g kg^{-1}	7089	0.00 to 33.30	2.57
Soil pH	Measure by 1:2.5 H_2O,	n/a	8266	2.30 to 9.59	6.16
BD	Soil bulk density	kg m^{-3}	4380	0.07 to 2.52	1.29
MBC	Soil microbial biomass carbon	mg kg^{-1}	4218	5.93 to 2986	413
Date	Observation month of DOC	month	12807	1 to 12	6.50
DOC_{phy}	Measure by physical method	mg kg^{-1}	3289	0.28 to 3181	155.99
DOC_{che}	Measure by chemical process	mg kg^{-1}	9518	0.04 to 7859	245.83
DOC	Soil dissolved organic carbon	mg kg^{-1}	12807	0.04 to 7859	222.78

https://doi.org/10.5194/essd-2024-343
Preprint. Discussion started: 29 August 2024

Open Access Earth System Science Data Discussions

Table 2. Global soil dissolved organic carbon concentration (mg kg^{-1}) for major ecosystems. 25% and 75% represent the 25th and 75th percentiles of one group, respectively. SD, Standard deviation; SE, Standard error.

Ecosystems	Mean	SD	SE	Skewness	Kurtosis	25%	Median	75%
Wetland	218.53	340.35	10.23	5.15	39.41	46.40	107.11	266.51
Forest	256.18	531.72	7.62	7.09	69.72	47.60	115.51	246.55
Shrubland	160.24	131.51	6.70	3.40	22.58	76.53	127.84	205.50
Tundra	470.78	721.70	63.30	4.67	29.59	86.91	241.09	577.00
Grassland	327.77	674.43	19.53	4.16	18.03	54.62	126.48	303.63
Cropland	165.98	272.51	3.81	6.53	73.25	40.51	83.00	178.81
Global	222.78	445.78	3.93	7.16	73.67	45.86	101.01	226.47

https://doi.org/10.5194/essd-2024-343
Preprint. Discussion started: 29 August 2024

Table 3. Analysis of the predicted global map of soil dissolved organic carbon. The area-weighted average soil dissolved organic carbon concentration was calculated based on our predicted map. Converting soil dissolved organic carbon concentration to soil dissolved organic carbon content and stock used the soil bulk density and land area.

Continent	Soil DOC concentration (mg kg^{-1})	Soil DOC content (g m^{-2})	Soil DOC stock (Pg)
Asia	274.43	107.79	4.80
North America	263.63	99.37	2.45
Europe	227.34	86.76	0.88
South America	215.81	77.05	1.37
Oceania	198.13	76.92	0.59
Africa	186.35	68.04	2.07
Global	237.56	89.80	12.17