Reviewer #1:

This study presents an innovative approach to deriving a high-resolution transformation rate (Pr) map of dissolved organic carbon (DOC) from soil organic carbon (SOC) across the contiguous United States using machine learning techniques, specifically XGBoost. By predicting DOC transformation rate based on various environmental attributes, the authors generate a DOC concentration reanalysis dataset for over two million small catchments. This research addresses the insufficient understanding of DOC conversion processes and provides a robust methodology for improving carbon cycle simulations in land surface and earth system models. The use of XGBoost to derive Pr values and create a high-resolution DOC concentration map is both novel and effective. The study's comprehensive data analysis and rigorous machine learning framework ensure robust and reliable results. Additionally, the findings have significant implications for enhancing carbon cycle models and informing climate change mitigation policies.

The article is logically structured, and the objectives are clear. The introduction is well-written and accessible, even to a geomorphologist like myself. According to the link provided in the 'Data availability' section, I downloaded all the Zenodo data. I opened and checked the raw data in ArcGIS, which includes SOC, PR, and DOC for the United States. The authors also provided a readme file explaining the attribute tables and their units.

If the authors can address my comments below and those of other reviewers through a major revision, the manuscript can be a valuable contribution to ESSD.

Response: We greatly appreciate the thorough and insightful review of our manuscript. We will carefully address the specific comments, which we believe will substantially improve our manuscript. In the following, the reviewer's comments, our point-to-point responses, and our proposed revision text are shown in black, blue, and purple colors, respectively.

Comments for manuscript:

Comment: This database provides an excellent method for calculating DOC. However, it is limited to the United States. If other scholars wish to apply this method to other regions, the contribution of this paper would be even greater. Although the authors mention in the first paragraph of the Method section that 'The methodology here is described with specific details over the CONUS region, but it is transferable to other regions after some modifications based on data availability', they do not provide further details on how to apply this method to other regions. Especially since some environmental factors have been found to have significant impacts while others do not. This information is crucial for applications in other regions. It is suggested that the authors discuss this briefly in the Potential use section (how the methods and results of this paper can inspire the calculation of DOC or Pr in other regions globally).

Response: Thank you for your valuable feedback and for highlighting the importance of extending our methodology to regions beyond the United States. We will add the following discussion regarding applying our methodology to other regions in the "Potential Use" Section:
"This methodology framework, in principle, can be extended to regions beyond CONUS. However, the specific implementations may vary from one region to another, depending on the availability and quality of the input data, particularly DOC observations and associated catchment attributes. When gathering catchment attributes, it is advisable to collect as many attributes as possible. During the selection process, it is recommended to avoid making independent assumptions and instead allow the model to determine important attributes. More Attention could be given to the attributes listed in Table 2 of this study, especially those with high feature importance rankings, such as the Subsurface flow contact time and

woody wetland percentage. Nonetheless, it is important to recognize that the representativeness and relative importance of attributes may vary by location."

Comment: The article is filled with numerous abbreviations, including those in the main text, figures, and tables, which increases the difficulty for readers. It is recommended that the authors avoid retaining so many abbreviations, especially those that appear infrequently in the later sections (e.g., fewer than five times) and those that are not used at all later on, for example, abbreviations like 'Pg' and 'ESMs' in the Introduction. In Tables 1, 2, 4, and 5, as a reader, I cannot understand what these parameters represent just by looking at the tables. Moreover, the parameter names in the 'Attributes' column of Table 1 are entirely unclear. Recommendations: Avoid using abbreviations in the tables. If abbreviations must be used, provide explanations at the bottom of the table to help readers understand, rather than having them search the main text or supplementary information for the full terms. The content in the 'Attributes' column of Table 1 is completely incomprehensible. It is suggested to move Table 1 to the supplementary information.

Response: Great point on the use of abbreviations in our paper. We will follow your suggestions and make the following changes:
1. We will replace the abbreviations that are infrequently used throughout the paper with their full spellings.
2. While the abbreviated attribute names are directly adopted from NHDPlus (hence, it is not practical to completely avoid using them), we will add explanations of each abbreviation in the footnote of each table and in the caption of each figure to enhance the readability.
3. We will move Table 1 to Supplementary to better streamline the main text.

Comment: The authors selected headwaters based on the following two criteria: 1) there are no upstream rivers flowing into them, and 2) their drainage areas are no more than 2500 km². I have the following questions:
1）Does the first criterion mean that the selected station can only have one river upstream without any tributaries?
2）If so, the second criterion excludes drainage areas larger than 2500 km². I find it hard to believe that a watershed of several thousand square kilometers has only one river without any tributaries. Please provide more details in the main text to clarify this and avoid confusion for readers like me.
3）Additionally, according to Fig. S1, the minimum drainage area in NHDPlus is 0.001 km², approximately a 30-meter square. As a geomorphologist, I do not understand how a watershed of this size can have sufficient upstream drainage area to form a river. Generally, river sources are not at the drainage divide but are 500-5000 meters downstream from it.
4）Moreover, I hope the authors clarify in the caption of Fig. S1 whether the data source includes all watersheds in NHDPlus or only the several thousand watersheds used in this study.

Response: Great questions. It appears that the confusion has been caused by our loose definition of "headwater catchment". We did not use 'headwater catchment' following its strict geomorphological definition based on stream order, river length, drainage area, or the number of tributaries (He et al., 2024). Rather, we used "headwater catchment" for the drainage area extending from the WQP station upstream to the furthest tributaries that do not have any upstream rivers. To avoid this type of confusion, we will use "small catchment" instead of "headwater catchment" from now on. Below, we also provide specific answers to each of your questions:

1) No, a small catchment can contain multiple tributaries.

2) It should be "small catchment". We will include the following definition of "small catchment" in Section 2.2 to prevent any confusion:
"In this study, a "small catchment" refers to the drainage basin extending from the WQP station upstream to the furthest tributaries that do not have any upstream rivers. Note that a small catchment is not necessarily a headwater catchment which includes only one river (He et al., 2024)."
3) Most of the NHDPlus local catchments are not headwater catchments but intermediate catchments, serving as connections between other catchments. For example, they may correspond to only a very small river segment (as part of a longer river), making their size quite small.
4) We will change the caption of Fig. S1 to the following:
"Figure S1. Distribution of the 2.6 million NHDPlus local catchment areas and flowline lengths."

Comment: The authors need to explain in the main text the format and size of all the files uploaded to Zenodo, especially detailing what information is included in the files with the format 'gpkg' and what software readers can use to open and edit them.

Response: We will add the following detailed information about all the files uploaded to Zenodo in the "Data Availability" Section:
"The Zenodo link contains four files: (1) **CONUS_PR_MAP.png**, a PNG image showing the model-simulated long-term averaged DOC transformation rate (Pr) across over 2.6 million NHDPlus local catchments; (2) **CONUS_DOC_MAP.png**, a PNG image depicting the long-term averaged DOC concentration reanalysis in soil leaching flux across over 2.6 million NHDPlus local catchments; (3) **DOC.gpkg**, a 9.9 GB GeoPackage file containing data on Pr and DOC for 2.6 million NHDPlus local catchments, including COMID, SOC concentration, and local catchment boundary polygons. This file is compatible with QGIS, ArcGIS, or Python libraries such as GeoPandas for opening and editing; and (4) **readme.txt**, a text file providing detailed information about the three aforementioned files."

Comment: Among the 29,320 WQP stations, some stations do not have existing upstream watershed boundaries. In such cases, the authors obtained the watershed boundaries using DEM. I have the following questions regarding this:
1) The authors need to clarify, among the 22,201 stations, how many watershed boundaries were derived from DEM and how many from NHDPlus?
2) What resolution of DEM was used, and how were the watershed boundaries calculated?
3) The authors should compare their calculated watershed boundaries with the global watershed boundaries based on a 90-m resolution DEM and advanced algorithms (ESSD 6, 1151–1166, 2024) and present a comparison figure in the SI.

Response: Thanks for these clarification questions. Below are our answers:
1) All watershed boundaries of the 22,201 stations were obtained using the HyRiver Python package (Chegini et al., 2021). This package does not directly derive catchment boundaries itself but retrieves them through The Hydro Network-Linked Data Index (NLDI) web server. Additionally, the package simplifies catchment boundaries and splits the catchment at the location of the WQP stations.
2) We did not use DEM to obtain watershed boundaries. We apologize for the incorrect information previously provided.
3) Since we did not derive the boundaries ourselves and the WQP stations can be located at any position along rivers (not necessarily at the outlet), it is not comparable with the global watershed boundaries based on a 90-m resolution DEM and advanced algorithms (ESSD 6, 1151–1166, 2024).
We will update the related paragraph as the following:

"We conduct a geospatial analysis to identify the upstream drainage area of each WQP river station using NHDPlus local catchments and flowlines. Utilizing the Python package HyRiver (Chegini et al., 2021), we co-located 29,320 WQP stations with the closest corresponding NHDPlus flowlines. However, 2,751 stations can not be linked due to the absence of adjacent flowlines. When WQP stations are in close proximity and share the same NHDPlus flowline, we retain only the station with the best data availability. For a given flowline, HyRiver traces back to every upstream flowline, accessing and merging the boundaries of all related NHDPlus local catchments (each flowline has its corresponding local catchment area) from the Hydro Network-Linked Data Index web server. It also requests the server to simplify boundaries and splits them precisely at the station locations. The relationship between the derived small catchment boundaries and the NHDPlus local catchments is shown in Figure S2a. Through this comprehensive geospatial analysis, we identify the upstream boundaries for 22,201 WQP stations."

Comment: The authors should provide a clear definition of "headwater" as used in this paper in the introduction. Is it determined based on stream order, river length, drainage area, or the number of tributaries? Additionally, they need to explain why they focus on headwaters.

Response: This comment has been essentially addressed in a previous response. We will use the terminology "small catchment" hereinafter, which refers to: a) DOC concentration at the outlet of a catchment is attributed to the entire upstream drainage area, and b) for small catchments, we can neglect the reaction of DOC in the stream at daily or sub-daily time steps, as explained in Section 2.1.

Comment: Maybe convert Table 2 into a bar chart and place it in SI. Additionally, again, many abbreviations in Table 2 need to be explained with their full terms.

Response: We will convert Table 2 into a bar chart, which we believe should remain in the main text as it highlights the importance of the selected features and is a critical result for Section 3.1. Additionally, the abbreviation issue has been addressed in a previous response by adding explanations to the figure captions and table footnotes.

Comment: The language in this article needs further refinement. Here are just some examples that need to be revised, and the authors should check the entire text:
1) Delete "quickly" from line 149.
2) Delete "required for this study" from line 153.
3) Refine "We collect a wide range of environmental variables, comprising a total of 126 variables" to "We collect 126 environmental variables.".
4) Change "The ML technique used in this study is the eXtreme Gradient Boosting (XGBoost) algorithm" to "We use the eXtreme Gradient Boosting (XGBoost) ML algorithm."

Response: We will carefully review and further refine the entire text. Regarding reviewer's specific comments:

1) We will delete "quickly" from line 149
2) We will delete "required for this study" from line 153
3) We will refine the original sentence to "We collect 126 environmental variables"
4) We will change the original sentence to "We use the eXtreme Gradient Boosting (XGBoost) ML algorithm"

Comment: The title is a bit long; it is recommended to change it to: "U.S. Transformation Rate Map of Dissolved Organic Carbon" or "Transformation Rate Map of Dissolved Organic Carbon in the Contiguous U.S."

Response: Great suggestion. We will change the title to "Transformation Rate Map of Dissolved Organic Carbon in the Contiguous U.S." to enhance clarity and conciseness.

Comment: The citation format for figures is completely inconsistent throughout the text. Examples for the same figure include: Fig. S1, supplementary Fig. S1, and Supplementary Fig. S1. Please check the entire text (main text, figures, SI) and standardize according to ESSD requirements.

Response: We will thoroughly review the entire manuscript, including the main text, figures, and supplementary information, to ensure that all figure citations are standardized according to the ESSD requirements. Specifically, we will consistently use the format "Fig. S1" throughout the manuscript.

Comment: L175 ScienceBase also provides indicators of human activities, right?

Response: Yes, you are correct. ScienceBase provides catchment attributes across 11 categories, including human activities. In our study, we present the four categories that we found significant in predicting DOC. We will refine the sentence as follows to avoid confusion:
"ScienceBase includes a wide range of environmental variables across 11 categories, such as climate, hydrology, soil, and geological data, conveniently available at the catchment scale across the entire CONUS."

Comment: L244 "Out of the remaining 95 variables (see supplementary Tables S1 and S2 for details), 46 are relatively independent from each other. However, the other 49 are highly correlated with one or more variables." How did the authors determine "relatively independent" and "highly correlated"? I expect to see more explanation of this in the main text.

Response: Thanks for pointing out this, we will add the following clarification:
"A "correlated group" is defined as a set of variables that are highly correlated, with a Pearson correlation coefficient ≥ 0.8 or ≤ -0.8. Within this group, every variable has at least one other variable that it is highly correlated with, and not with those outside the group. A variable is considered relatively independent if its correlation coefficient with all other variables is < 0.8 or > -0.8. The correlation threshold of ± 0.8 is adopted following the guidelines by Schober et al. (2018)."

Comment: Line 249, change "see Supplementary Figure S3" to "Supplementary Figure S3." Please check the entire text for similar instances where "see" is unnecessary.

Response: We will change "see Supplementary Figure S3" to "Supplementary Figure S3" on line 249. Additionally, we will carefully review the entire text to identify and remove any unnecessary instances of "see" in figure citations to enhance clarity and consistency.

Comment: Line 251: "This new variable is thus independent of the other environmental variables." I do not understand the basis for this statement. Even if the new 9 combined parameters are formed, it is unlikely that they are completely independent of the other 46 parameters. The authors should provide a brief explanation in the main text or delete this sentence.

Response: Our original intent was to convey that after this treatment, the newly merged, single variable is considered to be relatively independent of the other environmental variables. We will revise the sentence as follows:
"This new variable is now relatively independent of the other environmental variables."

Comment: Lines 273-275 need to be supported by references.

Response: Those two sentences, "Recent studies have demonstrated the efficiency and effectiveness of these techniques in capturing high-dimensional and complex relationships between a target biogeochemical variable and various environmental predictors," and "These techniques have been successfully applied in various studies, including riverine sediment, beach water quality, oceanic particulate organic carbon, and eutrophication impacts from corn production (Abeshu et al., 2022; Li et al., 2022; Liu et al., 2021; Romeiko et al., 2020; Fan et al., 2021)," share the same citations. We will modify these two sentences as follows:
"These techniques have been successfully applied in various studies, including riverine sediment, beach water quality, oceanic particulate organic carbon, and eutrophication impacts from corn production (Abeshu et al., 2022; Li et al., 2022; Liu et al., 2021; Romeiko et al., 2020; Fan et al., 2021), demonstrating their efficiency and effectiveness in capturing high-dimensional and complex relationships between a target biogeochemical variable and various environmental predictors."

Comment: Line 379: "per_canopy" is too difficult to understand.

Response: This has been addressed in a previous comment. All the abbreviations in the "Name used in this study" column of Table 2 will be deleted. Instead, we will use the original names as provided in the NHDPlus dataset.

Comment: In some places, it is written as "section," while in others, it is abbreviated as "sect" (e.g., L380).

Response: Thanks. We will review the entire manuscript to ensure consistent usage of the term "section."

Comment: L413' Note the unit of DOC concentration in water is mostly reported in mg/L (Schelker et al., 2012; Tian et al., 2015b; Langeveld et al., 2020)'. I think this sentence is not important to be in the main text.

Response: Agree, it will be deleted from the main text.

Comment: L481-482 "Blue, red, and grey colors are employed to indicate whether dropping the corresponding predictor will result in an increase, decrease, or insignificant change in the model's performance, respectively" should be in figure caption, rather than here.

Response: We will move the explanation about the colors indicating changes in model performance from lines 481-482 to the caption in Figure 7.

Comments for dataset in Zenodo:

Comment: There are many blank "nodata" areas within the CONUS_DOC_MAP, whereas the CONUS_PR_MAP does not have this issue. The authors need to explain this in the main text.

Response: It has already been explained in the first paragraph of section 5 as: "Due to missing data in the HWSD 1km SOC map at about 0.6 million NHDPlus local catchments, we cannot calculate the C_{DOC_runoff} values over those catchments."

Comment: For reproducibility, the authors need to provide the shapefiles (or other similar vector data) for the 2595 watersheds used for machine learning training and the 3210 watersheds used for evaluation, as well as the shapefiles for these 5805 stations. The machine learning codes, as well as the raw data used for training the machine learning model, need to be uploaded to Zenodo; Then provide another link in the manuscript (not https://doi.org/10.5281/zenodo.8339372).

Response: Thank you. We will upload the shapefiles for the 2,595 watersheds used for machine learning training, the 3,210 watersheds used for evaluation, the 5,805 stations, the machine learning codes, and the raw data used for training the model to Zenodo. We will provide the new Zenodo link in the manuscript to ensure reproducibility.

Suggestion for figures:

Comment: The background color of all 2D density plots needs to be changed because the background color is included in the color scale. This makes it difficult for readers to distinguish between the data and the background color.

Response: Good catch! We will change the background color of all 2D density plots to light grey to ensure it is not included in the color scale.

Comment: Are the points in Figure 1 outlets or geometric centers of the watersheds? Additionally, it is necessary to indicate in the figure or caption that the gray lines represent rivers and the black lines represent national boundaries. Also, please specify the sources of these two elements.

Response: Thank you for the question. The points represent the locations of the WQP stations, which are also the outlets of the corresponding small catchments. The CONUS boundary was obtained from the GeoPandas built-in shapefile data, accessible through geopandas.read_file(gpd.datasets.get_path('naturalearth_lowres')).
The river shapefile was obtained from Natural Earth. We believe it is not very important to mention the source in the main text, as both are open-source data. We will revise the caption as follows:
"The points indicate the locations of the WQP stations, which are also the outlets of the corresponding small catchments. The CONUS boundary and river shapefiles are directly from open-source datasets GeoPandas (Van den Bossche et al., 2024 and Natural Earth (Made with Natural Earth. Free vector and raster map data @ naturalearthdata.com.), respectively."

Comment: Figures 4 and 7 contain numerous abbreviations that are not explained in the captions, making it difficult for readers to understand the figures directly.

Response: This issue has been addressed previously by adding explanations into the captions.

Comment: It is necessary to explain in the caption of Figure 4 what the correlation coefficient is. Is it Spearman rank?
Response: We will revise the caption of Figure 4 to specify that the Pearson correlation coefficient is used. The updated caption now reads:
"Figure 4. Covariance heatmap of Pr and the 12 selected NHDPlus predictors. The Pearson correlation coefficient is used."

Comment: Why are there many nodata areas near the national boundaries in Figure 5?

Response: Nice catch! This discrepancy is not due to missing data but rather a mismatch between two geo-datasets. The country boundaries, obtained from gpd.datasets.get_path("naturalearth_lowres") in the GeoPandas library, are of very low resolution. In contrast, the NHDPlus local catchments are derived from a 30m DEM, which are much more accurate. This difference in resolution leads to discrepancies at the national boundaries. Additionally, for approximately 14,000 NHDPlus local catchments, we cannot retrieve their catchment attributes, so they are removed from the prediction set.

Comment: Fig. S1 needs ticks on the X-axis.

Response: We will add ticks on the X-axis of Fig. S1.

Comment: In the main manuscript, I do not understand the differences between the two types of watershed boundaries provided by NHDPlus. Besides, I do not understand Figure S2. It is recommended to use a real terrain example for illustration to show the differences between these two kinds of watershed boundaries. For example, based on Google Earth, mark the river, the two different watersheds, and the DOC station location (outlet).

Response: Upon rechecking the NHDPlus version 2.1 National Seamless Geodatabase (.gdb), we found that it contains only the boundaries of local catchments. We apologize for the incorrect information provided earlier, and will update the corresponding paragraphs in the main text. To enhance clarity, we will regenerate Figure S2 to include two subplots: a) demonstrating the relationship between small catchments and the NHDPlus local catchments within them, and b) illustrating the nesting of small catchments and how we handle them.

Comment: I don't have research experience with DOC; most of my comments are from a geomorphological perspective, as well as regarding readability and clarity. I hope my suggestions are helpful.

Response: Yes, your suggestions are indeed very helpful for improving the readability and clarity. We are very grateful.