Review report for ESSD-2024-43

'Deriving a Transformation Rate Map of Dissolved Organic Carbon over the Contiguous U.S.'

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.

Comments for manuscript:

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).

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.

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.

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.

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.

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.

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.

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."

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."

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.

L175 ScienceBase also provides indicators of human activities, right?

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.

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

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.

Lines 273-275 need to be supported by references.

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

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

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.

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.

Comments for dataset in Zenodo:
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.

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).

Suggestion for figures:
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.

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.

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

It is necessary to explain in the caption of Figure 4 what the correlation coefficient is. Is it Spearman rank?

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

Fig. S1 needs ticks on the X-axis.

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).

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.

Chuanqi He MIT, USA 21 May 2024