https://doi.org/10.5194/essd-2024-43
Preprint. Discussion started: 2 April 2024

Open Access Earth System Science Data Discussions

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

Lingbo Li[1], Hong-Yi Li [1*], Guta Abeshu[2], Jinyun Tang[3], L. Ruby Leung[2], Chang Liao[2], Zeli Tan[2], Hanqin Tian[4], Peter Thornton[5], Xiaojuan Yang[5]

[1]Department of Civil and Environmental Engineering, University of Houston, Texas, USA
[2]Pacific Northwest National Laboratory, Washington, USA
[3]Lawrence Berkeley National Laboratory, California, USA
[4]Boston College, Massachusetts, USA
[5]Environmental Sciences Division, and Climate Change Science Institute, Oak Ridge National Laboratory, Tennessee, USA

Correspondence to: Hong-Yi Li (hongyili.jadison@gmail.com)

Abstract. Riverine dissolved organic carbon (DOC) plays a vital role in regional and global carbon cycles. However, the processes of DOC conversion from soil organic carbon (SOC) and leaching into rivers are insufficiently understood, inconsistently represented, and poorly parameterized, particularly in land surface and earth system models. As a first attempt to fill this gap, we propose a generic formula that directly connects SOC concentration with DOC concentration in headwater streams, where a single parameter, the transformation rate from SOC in the soil to DOC leaching flux, P_r, accounts for the overall processes governing SOC conversion to DOC and leaching from soils (along with runoff) into headwater streams. We then derive a high-resolution P_r map over the contiguous U.S. (CONUS) in five major steps: 1) selecting 2595 headwater catchments where observed riverine DOC data are available with reasonable quality; 2) estimating catchment-average SOC for the 2595 catchments based on high-resolution SOC data; 3) estimating the P_r values for these catchments based on the generic formula and catchment-average SOC; 4) developing a predictive model of P_r with machine learning (ML) techniques and catchment-scale climate, hydrology, geology, and other attributes; and 5) deriving a national map of P_r, based on the ML model. For evaluation, we compare the DOC concentration derived using the P_r map and the observed DOC concentration values at another 3210 headwater gauges. The resulting mean absolute scaled error and coefficient of determination are 0.73 and 0.47, respectively, suggesting the effectiveness of the overall methodology. Efforts to constrain uncertainty and evaluate sensitivity of P_r to different factors are discussed. To illustrate the use of such a map, we derive a riverine DOC concentration reanalysis dataset for more than two million small catchments over CONUS. The P_r map, robustly derived and empirically validated, lays a critical cornerstone for better simulating the terrestrial carbon cycle in land surface and earth system models. Our findings not only set a foundation for improving our predictive understanding of the terrestrial carbon cycle at the regional and global scales but also hold promises for informing policy decisions related to decarbonization and climate change mitigation.

Notice: This manuscript has been authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (https://www.energy.gov/doe-public-access-plan).

1 Introduction

With the Earth's climate rapidly warming due to increasing atmospheric greenhouse gas concentrations, there is a growing focus on quantifying the regional and global carbon pools within the land, riverine, and oceanic systems, as well as the intricate interconnections among them (Jing et al., 2021; Teodoru et al., 2015; Duarte, 2017). Each year, about 2 billion metric tons (Pg) of dissolved organic carbon (DOC) are transported from land to the oceans via rivers globally, comparable to the amount of atmospheric CO2 that deposits into the ocean (Hansell et al., 2009; Lønborg et al., 2020). Moreover, riverine DOC is vital to aquatic biogeochemistry by providing nutrients to microbial communities and influencing aquatic greenhouse gas emissions (Li et al., 2019).

However, it remains a challenge to represent and predict riverine DOC effectively in the land biogeochemical module of Earth system models (ESMs), which are the primary tools for studying carbon cycles in the context of climate change. A chief reason behind this long-standing challenge is the complexity of terrestrial and aquatic processes and their interactions governing SOC transformation to DOC and transport from soils to rivers. The relevant terrestrial processes include the conversion of solid SOC into soil DOC, the adsorption and desorption of DOC by surrounding soils, the transport of DOC from soils into headwater streams along with runoff, and the degradation of soil DOC during this transport. The relevant aquatic processes include the transportation of riverine DOC from headwater streams, the interception of DOC fluxes by reservoirs and lakes, the degradation of riverine DOC during transport, and the consumption of DOC by aquatic biosystems. Furthermore, each process is controlled by several environmental factors, which often exhibit substantial spatial heterogeneity. Models attempt to represent these complexities through parameters associated with governing equations. For instance, Tian et al. (2015a, b) incorporated the effects of runoff on DOC leaching with a coefficient that involves both surface and subsurface runoff. Surface and subsurface runoff are further affected by many environmental factors such as climate, soil, vegetation, and topography (Li et al., 2014; Li and Sivapalan, 2014).

The complexity of relevant processes and their driving environmental factors is also evident in the diverse process descriptions in several land biogeochemical models that are pioneers in representing the suite of processes from SOC to riverine DOC, such as Dynamic Land Ecosystem Model (DLEM) (Tian et al., 2015a, b; Yao et al., 2021), the integrated catchment model for carbon (INCA-C) (Futter et al., 2007), the Joint UK Land Environment Simulator Dissolved Organic Carbon model (JULES-DOCM) (Nakhavali et al., 2018), and the TRIPLEX-hydrological routing algorithm (TRIPLEX-HYDRA) (Li et al., 2019). These models differ in the processes involved and the process descriptions, owing to the inconsistent understanding of relevant processes among the modeling community. For instance, DLEM and TRIPLEX-HYDRA both adopt CENTURY-like (Parton et al., 1987; Metherell et al., 1993) formulas to estimate DOC leaching fluxes (Tian et al., 2015a, b; Yao et al., 2021; Li et al., 2019), but with notably different ways of incorporating both soil and water-related factors. For instance, TRIPLEX-HYDRA includes an empirical coefficient to account for soil absorption of SOC before its dissolution and DOC degradation in soils, which are not explicitly accounted for in DLEM. TRIPLEX-HYDRA incorporates hydrologic effects by directly using the water flow rate, whilst DLEM uses a dimensionless ratio to account for these effects. Equally important, the available observations have not been fully used for estimating or calibrating the numerous DOC-related parameters at the regional and larger scales in a spatially continuous yet variable fashion. Existing models usually calibrate several DOC-related parameters against DOC observations at a limited number of river gauges, leading to the issue of overparameterization, where multiple combinations of parameter values can achieve the same simulation results (Sivapalan, 2005). Moreover, the resulting parameters often poorly reflect the spatial heterogeneity of underlying processes and environmental factors due to the limited spatial coverage of DOC observations (Futter et al., 2007; Tian et al., 2015a, b; Nakhavali et al., 2018; Li et al., 2019; Liao et al., 2019; Yao et al., 2021). Overall, existing models for simulating DOC fluxes are still subject to limited transferability over poorly observed regions due to insufficient process understanding, data scarcity, and overparameterization.

One traditional strategy for improving model transferability over poorly observed regions is parameter regionalization. Generally, the low-dimensional relationships between a target parameter and other environmental variables are derived based on prior knowledge or regression analysis from the locations where sufficient observations are available. The relationships are then generalized and transferred to poorly-observed places (Doron et al., 2011; Dupas et al., 2013; Ye et al., 2014; Alebachew et al., 2014; Ayata et al., 2018; Tan et al., 2022). However, such a strategy will not work well if statistically robust and mechanistically meaningful relationships can not be derived from the conventional regression analyses or prior knowledge when, for example, the relationships are high-dimensional and nonlinear (Abeshu et al., 2022; Li et al., 2022). Fortunately, state-of-the-art machine learning (ML) techniques offer a promising and effective alternative strategy, owing to their proven advantages in capturing higher-order relationships between the target and predictive variables (predictors), especially when prior knowledge of such relationships is still in its infancy (Afan et al., 2016). For example, ML techniques have been successfully employed to capture the complex relationships among median sediment particle size (D50) and several environmental factors, which enabled the derivation of a national map of D50 (Abeshu et al., 2022). They have also been used to predict the concentration of fecal indicator bacteria, providing valuable guidance to beach closure problems (Li et al., 2022).

As the first step in addressing these challenges, this study develops an ML-powered approach for parameterizing DOC leaching fluxes at regional and continental scales. The rest of this paper is organized as follows. Section 2 outlines the overall methodology, including governing equations and corresponding parameters, data preparation, and the ML techniques employed. Section 3 presents the results over the contiguous United States (CONUS). Sections 4, 5, and 6 discuss the uncertainty, potential use of the resulting datasets, limitations of methods, and data availability. Section 7 concludes with a summary and potential future directions.

2 Methods

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.

2.1 Governing Equation

Several existing land or land biogeochemical models commonly employ CENTURY-like formulas to represent the leaching of DOC (Futter et al., 2007; Tian et al., 2015a, b; Nakhavali et al., 2018; Li et al., 2019; Yao et al., 2021; Parton et al., 1998). In such formulas, the DOC leaching flux is estimated as a linear function of several factors, including the SOC or DOC concentration in soil, runoff, and other relevant environmental factors. For example, in DLEM (Tian et al., 2015a, b), DOC leaching flux is estimated as

$$F_{DOC\_runoff} = F_{SOC\_Soil} \times \alpha 1 \times \alpha 2 \times \alpha 3 \qquad (1)$$

Where F_{SOC_Soil} is the total amount of decomposed SOC in soil (g Cm^{-2}s^{-1}); $\alpha 1$ is the fraction of decomposed SOC that is dissolvable (%); $\alpha 2$ is the runoff coefficient (-), i.e., the ratio of total runoff volume to the sum of total runoff volume and soil water content; and $\alpha 3$ is another coefficient (-) accounting for the effects of DOC concentration in soil water and desorption. In TRIPLEX-HYDRA (Li et al., 2019), DOC leaching flux is given as

$$F_{DOC\_runoff} = C_{SOC} \times K_s \times K_a \times Q_{runoff} - K_{soil} \qquad (2)$$

where F_{DOC_runoff} is the DOC flux in the soil water (g C/s); C_{SOC} is the concentration of SOC in the soil (g C/m^3); K_s is the solubility of SOC (-); K_a is the adsorption coefficient of SOC (-); K_{soil} represents the degradation rate of DOC in soils (g C/s), and Q_{runoff} is total runoff rate (m^3/s).

Based on the similarity between equations (1) and (2), while keeping minimal complexity in the process representation, we propose a simpler formula to estimate DOC leaching flux as

$$F_{DOC\_runoff} = C_{SOC} \times Q_{runoff} \times P_r \qquad (3)$$

Eqn. (3) can be rewritten as

$$C_{DOC\_runoff} = \frac{F_{DOC_{runoff}}}{Q_{runoff}} = C_{SOC} \times P_r \qquad (4)$$

Where F_{DOC_runoff} is the DOC leaching flux (g C/s), C_{SOC} is the SOC concentration (g C/m^3 soil), Q_{runoff} is the runoff volume per unit time (m^3 water/s), P_r is the transformation rate from SOC in soil to DOC in runoff (m^3 soil/ m^3 water), and C_{DOC_runoff} is the DOC concentration in the runoff (g C/m^3 water).

Eqn. (4) has several advantages: 1) its lumped parameter, P_r, accounts for all relevant processes and factors, including soil carbon decomposition, DOC sorption-desorption balance, DOC transport and degradation in soils, etc.; 2) its simplicity significantly reduces data requirements for large-scale parameterization since it is highly parameter-parsimonious and much more compatible with the availability of DOC observational data.

We further assume that C_{DOC_runoff} can be approximated with the riverine DOC concentration at the catchment outlets for headwater catchments, i.e.

$$C_{DOC\_outlet} \approx C_{DOC\_runoff} \tag{5}$$

Where C_{DOC_outlet} is the riverine DOC concentration at the catchment outlet (g C/m^3). The rationale behind Eqn. (5) is two-fold: 1) The travel time of runoff in small headwater streams is typically much less than one day, e.g., the daily total runoff rate can be approximated with the daily streamflow rate for headwater catchments (Li et al., 2013; Ducharne et al., 2003), and 2) Due to the short travel time of DOC in headwater streams, riverine DOC degradation in headwater streams mostly occurs at a rate of about 1% per day according to previous experimental and modeling studies (Strauss & Lamberti, 2002; Tian et al., 2015a,b; Li et al., 2019), hence is negligible.

Combining Eqn. (4) and (5) yields

$$C_{DOC\_outlet} \approx C_{SOC} \times P_r \tag{6}$$

Eqn. (6) may be used in at least two ways: 1) One can estimate P_r at the catchment scale wherever observed DOC concentration and SOC values are available, and 2) Once P_r is estimated a priori or through calibration, one can quickly predict riverine DOC concentration or discharge in headwater streams from the corresponding SOC values.

2.2 Data

A key step in the data preparation in this study is to pair up SOC data and riverine DOC observations at headwater catchments. The SOC data required for this study are from the Harmonized World Soil Database (HWSD) v1.2 (Fischer et al., 2008). This database provides SOC values at a spatial resolution of 1 km for two vertical soil layers at each grid cell - the top layer (0-30 cm) and the sub-layer (30–100 cm). Considering that DOC leaching from soils into rivers predominantly comes from the topsoil (Brooks et al., 1999; Finlay et al., 2006), we use the SOC content data from the top 30 cm layer for our estimations. We also take into consideration that there are missing values in some grid cells in the HWSD v1.2 and adjust our catchment selection accordingly. Riverine DOC observations are available via the Water Quality Portal (WQP) (Water Quality Portal, 2021). WQP integrates the publicly available water quality data from the USGS National Water Information System (NWIS) (U.S. Geological Survey), the EPA STOrage and RETrieval Water Quality eXchange (STORET-WQX) (USEPA), and the USDA ARS Sustaining The Earth's Watersheds - Agricultural Research Database System (STEWARDS) (Steiner et al., 2008). As of now, the WQP features data from 32071 river stations within the CONUS. These stations have recorded at least one DOC measurement between 1900 and the present.

In order to pair up SOC and DOC data at headwater catchments, we rely on the National Hydrography Dataset Plus (NHDPlus) dataset hosted by the U.S. Geological Survey (USGS) (Mckay et al., 2012). This dataset is chosen for two reasons: Firstly, NHDPlus provides well-defined catchment boundaries and their corresponding river segments, denoted as flowlines. There are ~2.6 million NHDPlus flowlines in CONUS, each with its corresponding local catchment boundary and other environmental attributes. For each flowline, there are two types of catchment boundaries provided: a local catchment which is immediately adjacent to and collects lateral runoff into the flowline, and an upstream drainage catchment which is the sum of both local catchment and the drainage areas corresponding to all the flowlines upstream of the local one. The sizes of these 2.6 million local catchments vary from the 5th percentile at 9.68 km^2 to the 95th percentile at 0.02 km^2, depending on the corresponding surface topography, with a CONUS average of 3.12 km^2 (see supplementary Figure S1). Secondly, NHDPlus is closely linked to ScienceBase (Wieczorek et al., 2018), a comprehensive scientific data and information management platform also hosted by USGS. ScienceBase incorporates a wide range of environmental variables, including climate, hydrology, soil, and geological data, conveniently available at the catchment scale over the whole CONUS. These environmental data are critical in the ML modeling analysis.

Correspondingly, the overall data preparation procedure consists of three major steps: 1) Selection of headwater catchments based on the availability of observed riverine DOC concentrations of adequate quality. 2) Estimation of P_r values for the catchments selected in Step 1, leveraging the corresponding riverine DOC observations and SOC reanalysis data. 3) Extraction of catchment-scale environmental variables that could potentially influence P_r. Specific details of each step will be further discussed in the following subsections.

2.2.1 Selecting headwater catchments

Our selection process for suitable headwater catchments involves the integration of the NHDPlus dataset and observed riverine DOC concentration data from river stations:

1. We conduct a geospatial analysis to identify the upstream drainage area of each WQP river station. This is accomplished by using the NHDPlus local catchments and flowlines. For every WQP station, we search for a NHDPlus flowline on which the station is located. Using a Python package HyRiver (Chegini et al., 2021), we co-locate 29320 WQP stations with the corresponding NHDPlus flowlines. However, the remaining 2751 stations cannot be linked with the NHDPlus dataset due to the absence of adjacent flowlines. Some WQP stations are in close

proximity to each other and share the same NHDPlus flowlines. In such a case, we retain only one WQP station with the best data availability. Each flowline in NHDPlus is accompanied by a corresponding watershed boundary. However, not all WQP stations are precisely located at the outlets of these existing NHDPlus watershed boundaries. When faced with these circumstances, we derive the upstream drainage area boundaries for the WQP stations from Digital Elevation Model (DEM) data. Upon completion of this comprehensive geospatial analysis, we identify the upstream boundaries for 22,201 WQP stations.

2. We further select the WQP stations whose drainage areas can be considered headwater catchments, based on two criteria: 1) there are no upstream rivers flowing into them, and 2) their drainage areas are no more than 2500 km^2. This size threshold ensures that the travel distance of river water (and consequently, DOC) is ~50 km within these catchments. Assuming an average channel velocity of ~1.0 m/s (Chow et al., 1988), the average travel time is ~14 hours, i.e., less than one day. Using these criteria, we identify 18,612 pairs of WQP stations and headwater catchments.
3. For the 18,612 WQP stations, we perform a rigorous DOC data quality control based on five criteria: a) The record lengths of riverine DOC data should span at least one year; b) There should be at least two riverine DOC observations; c) No single season should dominate the riverine DOC observations, i.e., a single season should not account for more than 50% of the records; d) within the boundaries of the corresponding catchments, there should be sufficient availability of the NHDPlus catchment attributes and SOC reanalysis data; e) the catchments should not be significantly affected by dams, i.e., the total drainage areas of the dams within a catchment should be no more than 5% of the total catchment area. The adoption of criteria (a)-(e) reflects a careful balance between ensuring data quality and maintaining adequate quantity, ensuring that sufficient WQP stations are retained to represent the entire CONUS. After the data quality control, there remain 5805 WQP stations with their corresponding headwater catchments.
4. For the 5805 WQP stations and their headwater catchments, we verify the spatial independence among them. For instance, Catchment A is considered to be nested within Catchment B if A is situated within the drainage area of B. In such scenarios, while the fluxes observed at the outlet of Catchment B are dependent on those at the outlet of Catchment A, Catchment A itself remains independent of B. As illustrated in Supplementary Figure S2, in cases of nested catchments, the catchment with the smaller area is consistently selected as the independent catchment. From the 5805 pairs of the WQP stations and catchments, we identify 2595 as being independent and suitable for further ML modeling. The other 3210 pairs, despite the nesting issue, are still valuable; they are thus kept for evaluation of estimated DOC (see Section 3.4).

https://doi.org/10.5194/essd-2024-43
Preprint. Discussion started: 2 April 2024

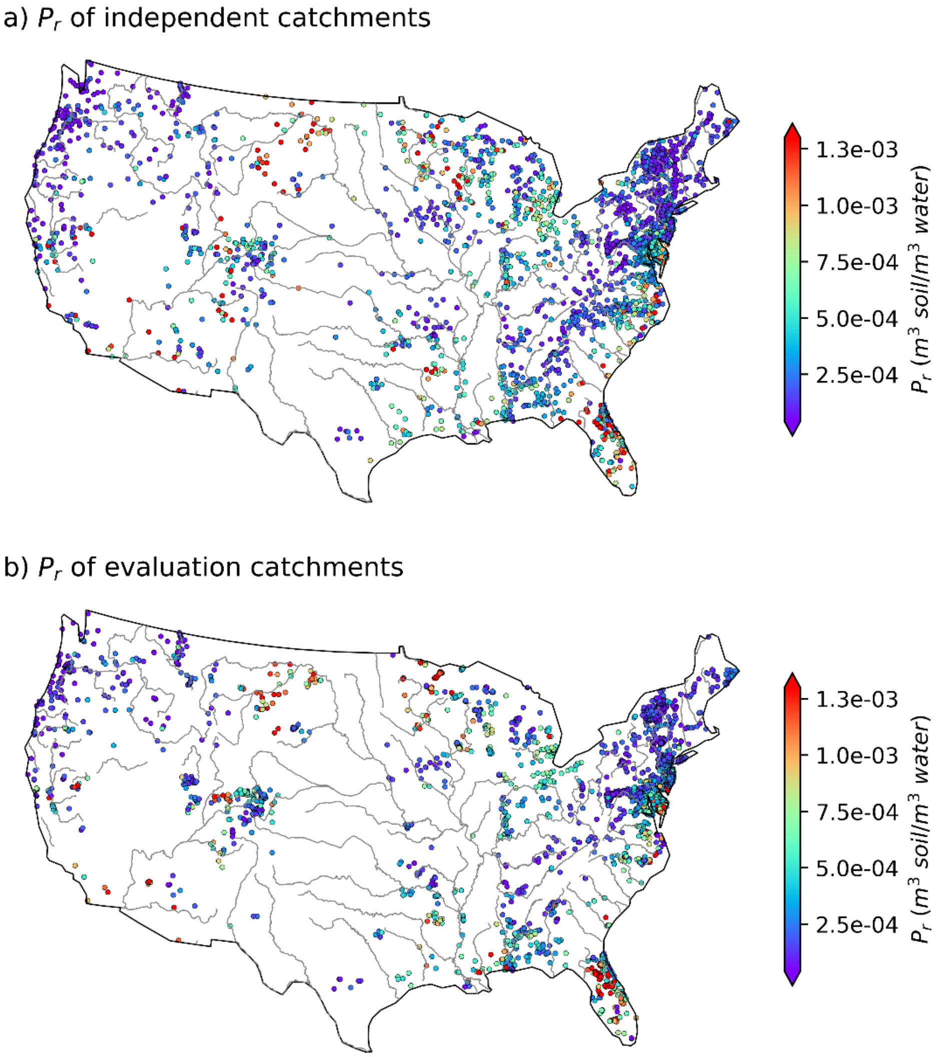
a) P_r of independent catchments
b) P_r of evaluation catchments
1.3e-03
1.0e-03
7.5e-04
5.0e-04
2.5e-04
P_r (m^3 soil/m^3 water)

Figure 1. Variability in estimated P_r across CONUS: a) For independent catchments (n=2595), and b) For evaluation catchments (n=3210). The color bars have been adjusted to enhance visual display by showing only the main body of values (from 5th percentile to 95th percentile)

2.2.2 Estimating P_r

For the final set of the paired WQP stations and headwater catchments, we calculate P_r using the DOC observation from the WQP stations and long-term mean SOC from HWSD based on Eqn. (6). For each catchment, the catchment polygons are used to clip the top-layer SOC map at the 1km resolution, and the catchment-scale SOC is subsequently calculated as the spatial average of SOC values at those 1km grid cells within the catchment. Hereafter the P_r estimated using Eqn. (6) are referred to as "*Estimated* P_r". The *Estimated* P_r, derived from the analysis of WQP DOC observations and HWSD SOC data, exhibits a wide range of values spanning several orders of magnitude. Figure 1a illustrates the spatial distribution of P_r for the 2595 independent catchments. In these catchments, the Estimated P_r ranges from 4.61x10^{-6} to 8.04x10^{-3} (m^3 soil/ m^3 water), with a median value of 2.50x10^{-4} (m^3 soil/ m^3 water). As a broad assessment of the similarity between the catchments used to construct the model and the evaluation catchments, values of P_r for the evaluation catchments calculated from data values of DOC and SOC using Eqn. (6) are shown in Figure 1b. Here, the *Estimated* P_r values in these catchments range from 8.81x10^{-6} to 6.37x10^{-3} (m^3 soil/ m^3 water), with a median of 2.60x10^{-4} (m^3 soil/ m^3 water). Note that the spatial distribution of the selected catchments is quite consistent with the spatial distribution of the WQP stations, i.e., more densely distributed in the eastern than western U.S, suggesting a good spatial representation of the selected catchments over all the WQP stations in CONUS.

2.2.3 Extracting environmental variables

The ScienceBase dataset is a comprehensive resource that houses a wide array of environmental variables sorted into categories such as climate, hydrology, geology, and land use/land cover. We collect a wide range of environmental variables, comprising a total of 126 variables, across eleven distinct categories. We remove seven attributes related to dams and streams from the analysis as they are irrelevant to our analysis objectives. Furthermore, we exclude 24 attributes from further analysis because they predominantly contain zero values, with over 80% of the values being zero over CONUS. 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. These 49 non-independent variables are further categorized into 9 "correlated groups" and named based on the group property, as listed in Table 1. A "correlated group" is characterized by interdependence within each "correlated group" in two steps. First, we normalize each variable within a group using the Yeo-Johnson power transformation (Yeo and Johnson, 2000) (see Supplementary Figure S3). The transformation ensures that the resulting dataset has a mean of 0.0 and a variance of 1.0. Second, we merge all the normalized variables into a single new variable through linear summation (Daoud, 2018). This new variable is thus independent of the other environmental variables. For those 46 independent variables, we apply the same transformation to minimize the impacts of varying magnitudes between different variables. Eventually, 54 independent variables remain, including 46 originally independent and 9 newly merged variables from the correlation groups.

Table 1. The 9 correlation groups and the corresponding merged NHDPlus attributes.

Correlation Group	Original NHDPlus Attributes
hydro_related	RECHG, WB5100_ANN, MAXP6190, PPT7100_ANN, RUN7100
temp_related	PET, FSTFZ6190, LSTFZ6190, PRSNOW, ET, TMAX7100, TAV7100_ANN, TMIN7100
agri_chem_related	FUNGICIDE, HERBICIDE, INSECTICIDE, N97, P97, NLCD01_82, PEST219, KGBI, KGCLADO, KGFISH
urban_related	POPDENS90, IMPV01_BUFF100, IMPV06, IMPV06_BUFF100, POPDENS00, POPDENS10, NLCD01_21, NLCD01_22, NLCD01_23, NLCD01_24, TOTAL_ROAD_DENS, HDENS10
soil_texture_related	SILTAVE, SANDAVE
soil_restrictive_related	SRL25AG, SRL35AG, SRL45AG, SRL55AG
wetd_related	MAXWD6190, WDANN
topo_related	EWT, TWI, BASIN_SLOPE

elev_related	ELEV_MEAN, ELEV_MIN, ELEV_MAX

2.3 Machine learning techniques

The ML technique used in this study is the eXtreme Gradient Boosting (XGBoost) algorithm, which is a powerful and widely adopted machine learning algorithm due to its exceptional performance in various applications (Abeshu et al., 2022; Delavar et al., 2019; Li et al., 2022). XGBoost is a scalable end-to-end tree-boosting system that belongs to the ensemble learning family (Chen and Guestrin, 2016). It combines multiple weak learners into a strong learner via sequential training and improving, and eventually forms a robust and accurate predictive model. By using XGBoost in this study, we aim to develop a predictive model that establishes causal linkages between the target variable, P_r, and a small number of environmental variables (denoted as predictors hereafter).

In addition to XGBoost, we take advantage of some other ML tools and techniques. Specifically, we use the Optuna optimization framework (Akiba et al., 2019) and k-fold cross-validation (k=5) for tuning the hyperparameters. By leveraging Optuna and k-fold cross-validation, we can systematically search and optimize the hyperparameters, maximizing the model's performance and accuracy. Furthermore, we employ the SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) to aid in the selection of environmental factors that are related to P_r. SHAP is a technique that assigns importance values to individual predictors in a model, providing insights into their contributions to the prediction. By using SHAP, we can identify the key environmental factors that significantly influence P_r and further refine our model. 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. 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). Readers are referred to Abeshu et al., (2022) for more details about these techniques.

The overall procedure for developing a predictive ML model is illustrated in Figure 2 and outlined as follows:

1. Prepare the input data for the ML modelling based on the independent catchments, their corresponding P_r estimates, and environmental variables. To address the substantial statistical disparities and wide variation within each predictor, we employ power transformation on all predictors. The lambda parameter is held constant during the transformation process for the training, testing, and prediction datasets to ensure consistent and reproducible results. Following the transformation, the dataset exhibits a zero-mean and unit variance, with a distribution that closely resembles a Gaussian distribution (as illustrated in Figure S3).
2. Randomly split the observational dataset (2595 catchments) into two sets: 70% for training and 30% for testing the ML model. These training and testing sets will be used throughout the subsequent steps.

3. Identify the list of predictors out of the 54 environmental variables extracted in Section 2.2.3 in three sub-steps:
    a. Generate a completely random predictor.
    b. Prepare an initial list of candidate predictors consisting of the random predictor and an initial list of candidate environmental variables. Use Optuna and k-fold cross-validation to obtain the optimal hyperparameters and train an intermediate ML model until the model achieves the best performance evaluated using the testing set.
    c. Calculate and rank the SHAP values for all the candidate predictors. Update the list of candidate predictors by keeping only those predictors with better SHAP values than the random predictor. For example, if the random predictor is ranked 20th, only the top 19 predictors are passed to the next iteration.
    d. Obtain an almost-final list of predictors by repeating sub-steps b-c.
4. Check the representativeness of the almost-final list of predictors identified in Step 3. For each of these predictors, check whether its values from the independent catchments are statistically representative of the whole CONUS, i.e., its values from those 2.6 million local catchments. Drop those predictors that cannot pass the representativeness check. Similar to Abeshu et al. (2022), the representativeness check on each of the almost-final predictors is performed by comparing the cumulative distribution function (CDF) derived from the observational dataset (2595 training catchments) and the CDF derived from the whole CONUS (about 2.6 million local catchments in NHDPlus). Specifically, comparisons are made between the 5th, 25th, 50th, 75th, and 95th percentiles between the two CDFs. After this Step 4, a final list of predictors is obtained.
5. Develop the final ML model based on the final list of predictors using Optuna and k-fold cross-validation methods.

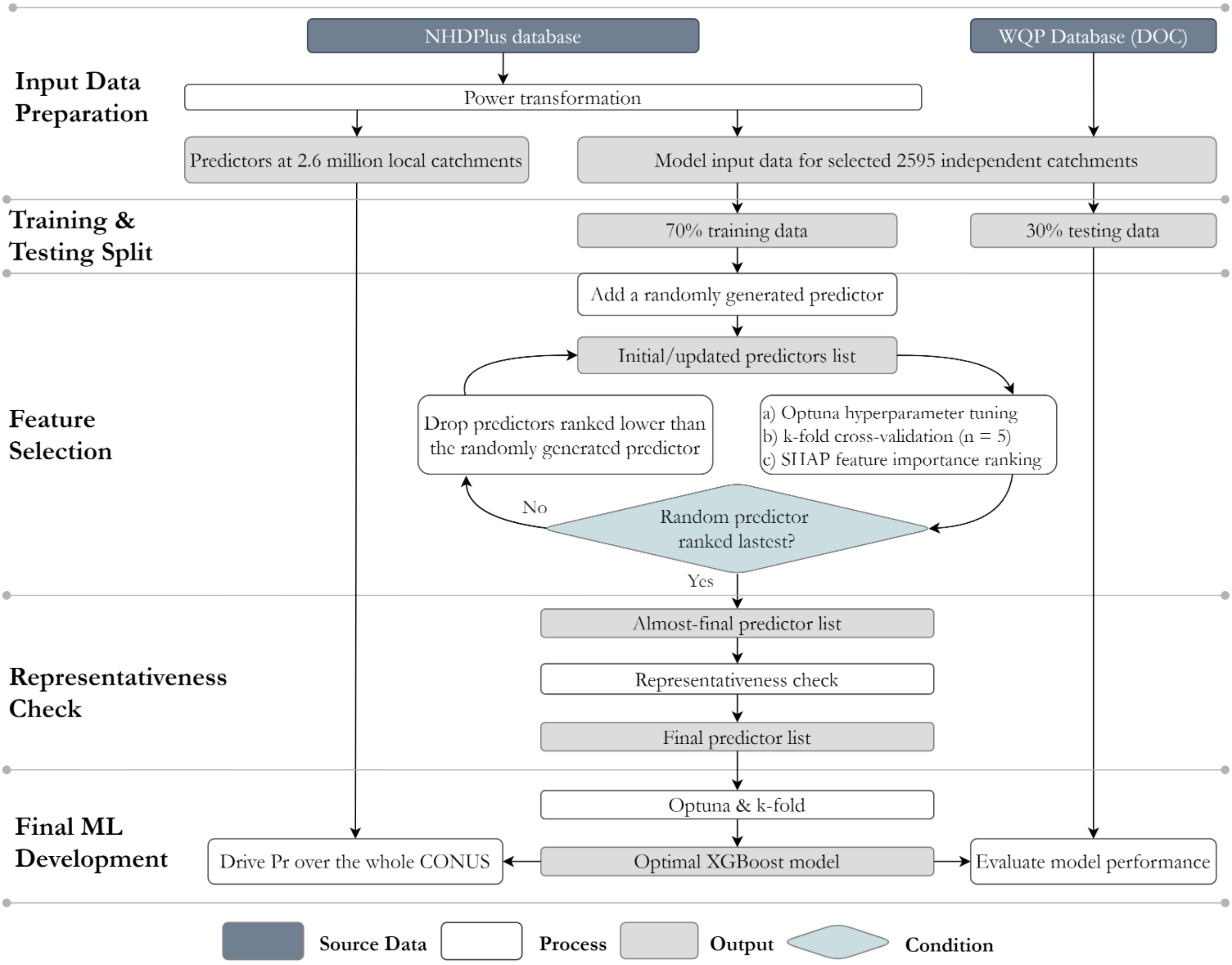
NHDPlus database
WQP Database (DOC)
Input Data Preparation
Power transformation
Predictors at 2.6 million local catchments
Model input data for selected 2595 independent catchments
Training & Testing Split
70% training data
30% testing data
Add a randomly generated predictor
Initial/updated predictors list
Feature Selection
Drop predictors ranked lower than the randomly generated predictor
a) Optuna hyperparameter tuning
b) k-fold cross-validation (n = 5)
c) SHAP feature importance ranking
No
Random predictor ranked lastest?
Yes
Almost-final predictor list
Representativeness Check
Representativeness check
Final predictor list
Optuna & k-fold
Final ML Development
Drive Pr over the whole CONUS
Optimal XGBoost model
Evaluate model performance
Source Data
Process
Output
Condition

Figure 2. A workflow for the XGBoost model.

In Steps 3 and 5, model performance metrics are required for model training and validation. The Kling-Gupta efficiency (KGE) (Gupta et al., 2009) has the advantage of simultaneously capturing both the magnitude and phase differences between the observed and simulated series (Gupta et al., 2009; Abeshu et al., 2022). However, further investigations have revealed several limitations: a) lack of an inherent benchmark value to distinguish between "good" and "bad" model performance, b) sensitivity to outliers, which can result in a systematic overestimation of the target variable, and c) instability when the target variable approaches zero (Pool et al., 2018; Santos et al., 2018; Knoben et al., 2019). Therefore, in addition to KGE, the mean absolute scaled error (MASE) is also used here to alleviate the influence of extreme values in the observation or simulation data (Hyndman and Koehler, 2006). MASE is a scaled error metric that is defined as the mean absolute error (MAE) of the model simulation divided by scaling factors (MAE of the observation in the original definition). In this study, we normalize MAE by the geometric mean of the observation data. Note that Steps 3 and 5 above are relatively independent of each other and do not have to rely on the same metrics.

3 Results

3.1 Predictor selection

In the predictor selection stage, after six iterations of hyperparameter tuning and predictor reduction with KGE as the metric, a list of 15 predictors is selected (see Table 2), including those related to climate, hydrology, pedology, and land cover. In addition, using MASE as the metric in this stage leads to a list of 19 remaining predictors, among which 13 are the same as the list of predictors identified using KGE. The predictor list selected using KGE is preferred due to the fewer predictors and similar model performance.

The most influential predictors, as determined by SHAP values, include the "hydro_related" group of hydrologic variables, the subsurface flow contact time ('index_tqsub'), the areal percentage of a soil class defined with a mixture of moderate and slow infiltration rates in a catchment ('per_soilmsI') (for more detailed definitions of soil classes, please refer to Ross et al., 2018), and the woody wetland percentage ('per_wwetland'). The "hydro_related" group of hydrologic variables is the linear summation of the annual average amount of runoff, precipitation, and groundwater recharge. Groundwater has a dilution effect on DOC concentration (Kortelainen and Karhu, 2006). Similarly, precipitation and runoff contribute to the distribution and concentration of DOC (Tranvik and Jansson, 2002; Baum et al., 2007; Wilson et al., 2013). The influence of subsurface flow contact time on DOC concentration is complex and indirect. For instance, during transport, a catchment with a shorter contact time experiences reduced mineralization loss (Ludwig et al., 1996) and microbial consumption (Helton et al., 2015). Conversely, studies have shown that labile DOC concentration increases with contact time in some alluvial aquifers, as deeper groundwater inflow could provide considerable labile DOC (Wickland et al., 2012; Helton et al., 2015). Soil type plays a crucial role in determining the soil organic matter quantity and the partitioning of precipitation into runoff, consequently influencing the concentration of DOC in rivers (Camino-Serrano et al., 2014; Autio et al., 2016). Woody wetland, as one land cover attribute, has been identified as a significant predictor of downstream DOC concentration (Duan et al., 2017), because of the enhanced breakdown of organic matter and plant respiration. To enhance the model transferability, a representativeness check (see Section 4.1.2) led to the exclusion of three predictors—'per_hwetland,' 'basin_area,' and 'per_shurb.' These variables, initially chosen, were found inadequate in representing the real-world data distribution anticipated during the prediction phase. Therefore, only 12 predictors are adopted in the final model training.

Table 2. Descriptions and SHAP values of 15 selected predictors

Predictor	Description

Name used in this study	Name in NHDPlus		Mean absolute SHAP value (15 attributes)	Mean absolute SHAP value (12 attributes)
hydro_related		Correlated group of hydrologic-related attributes	4.67E-05	5.96E-05
index_tqsub	CONTACT	Woody wetland percentage	3.80E-05	3.68E-05
per_soilmsI	HGBD	Areal percentage of Hydrologic Group B/D soil, with moderate infiltration rate when artificially drained and very slow infiltration rate while not drained	5.02E-05	5.37E-05
per_wwetland	NLCD01_90	Subsurface flow contact time index	3.52E-05	5.22E-05
ave_wetday	CWD	Average number of consecutive days with measurable precipitation	3.47E-05	3.78E-05
temp_related		Correlated group of temperature-related attributes	2.03E-05	3.42E-05
per_canopy	CNPY11_BUFF100	Percentage of tree canopy in 100-meter riparian buffer	1.83E-05	2.51E-05
elev_related		Correlated group of catchment elevation-related attributes	2.04E-05	2.30E-05
per_eforest	NLCD01_42	Evergreen Forest percentage	1.43E-05	1.72E-05
per_rhumidity	RH	Watershed average relative humidity percent	8.89E-06	1.43E-05
index_bflow	BFI	Base flow index	1.56E-05	1.37E-05
soil_texture_related		Correlated group of soil texture-related attributes	1.09E-05	1.09E-05
per_hwetland	NLCD01_95	Herbaceous wetland percentage	2.75E-05	
basin_area	BASIN_AREA	NHDPlusV2 flowline catchment area	9.37E-06	
per_shurb	NLCD01_52	Percentage of areas dominated by shrubs less than 5 meters tall	1.54E-05	

3.2 Final model

Figure 3 presents the performance of the ML model during both the training and testing phases (phases shown in Figure 2). To mitigate over-plotting, all the scatter plots (Figure 3 and hereinafter) employ color coding based on estimated density using kernel density estimation (KDE), as indicated by the corresponding color bar. After the exclusion of the three variables that displayed poor representativeness, the ML model performance remains stable between the training and testing phases, as gauged by metrics such as MASE, coefficient of determination (R^2), and normalized root-mean-square-error (NRMSE). The similarities in these metrics between the *Estimated* and predicted P_r values across both phases support the robustness of our 12-predictor model. Consequently, the final ML model and the subsequent analyses are based on the 12 selected predictors. Furthermore, the consistency of model performance between the training (MASE= 0.40) and testing (MASE= 0.81) phases suggests that the model overfitting issues are well-regulated (Ying, 2019). We also use KGE as the metric during the final model training. After a comparison between the modeling results using MASE (Figure 3) and KGE (supplementary Figure S4), MASE is preferred for two reasons: a) using MASE yields a better consistency in model performance between the training and testing phases, suggesting better model transferability; b) using MASE leads to a closer agreement between the model simulated and *Estimated* P_r values.

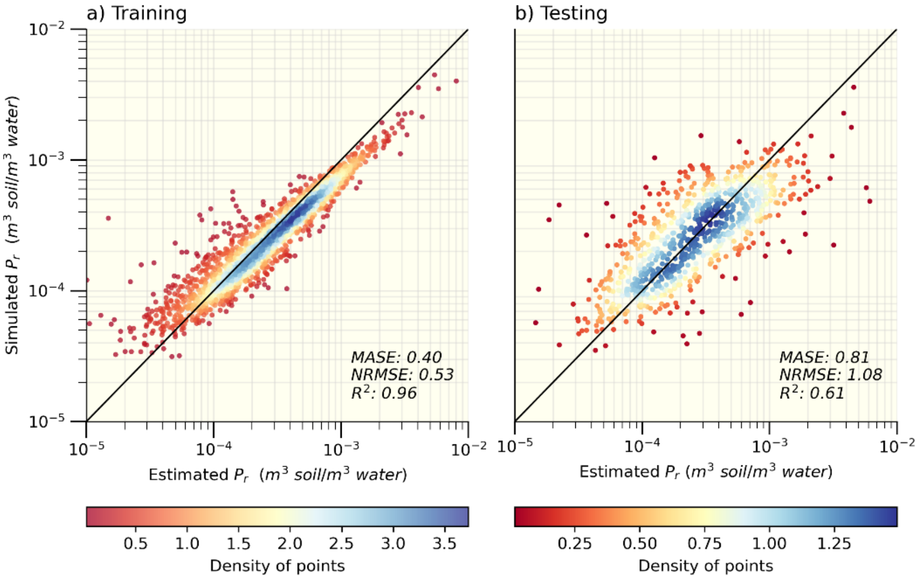

Figure 3. Performance of the XGBoost model with 12 predictors during a) the training phase (n=1816) and b) the testing phase (n=779). The solid black line indicates a 1:1 ratio. The varying colours indicate the density of points in the scatter plot.

Table 3 lists the optimized hyperparameter values of the final XGBoost model. We choose to tune 8 model parameters, which are critical to the XGBoost tree booster controlling regularization, subsampling, learning process, and the growth of the tree. The optimal values of model hyperparameters are quite different from the default ones, suggesting hyperparameter tuning is necessary.

Table 3. The optimal values of the XGBoost model hyperparameters.

Hyperparameter	Optimal Value	Tuning Range	Default value	Description
lambda	6.725×10^{-1}	[0, ∞]	1	Control L1 and L2 regularization; the larger the value, the more conservative the model will be
alpha	7.484×10^{-2}	[0, ∞]	0	
gamma	1.316×10^{-2}	[0, ∞]	0	Govern the model learning process by changing the step size shrinkage and minimum loss reduction; the larger the value, the more conservative the model will be
eta	1.277×10^{-1}	(0, 1]	0.3	
colsample_bytree	9.323×10^{-1}	(0, 1]	1	Control the subsample ratio of columns and training instances; a proper set of those values will prevent the model from over-fitting
subsample	6.142×10^{-1}	(0, 1]	1	
min_child_weight	8.410×10^{-2}	[0, ∞]	1	Determine the growth of the tree
Max_depth	12	[0, ∞]	6	

Figure 4 depicts the correlation between P_r and the 12 predictors and among the predictors themselves, where highly positive correlated and negative correlated are shown in dark-red and blue colors, respectively. Since we have treated the highly correlated variables, the highest positive correlation coefficient is 0.63 between "per_canopy" and "hydro_related", lower than the threshold of 0.8 we adopt in Sect 2.2.3. Among the observed correlation coefficients, the highest negative correlation coefficient, -0.69, is found between the variables "elev_related" and "temp_related." This strong negative correlation makes intuitive sense since air temperature decreases with increasing elevation. Note that all of the 12 selected predictors show weak or even negligible correlation with the target variable P_r, with the absolute values of the correlation coefficient less than 0.3. It is not surprising since the high-order, nonlinear relations between P_r and the predictors, and likely among the predictors themselves, can only be effectively captured by the ML techniques but not the traditional regression analysis methods.

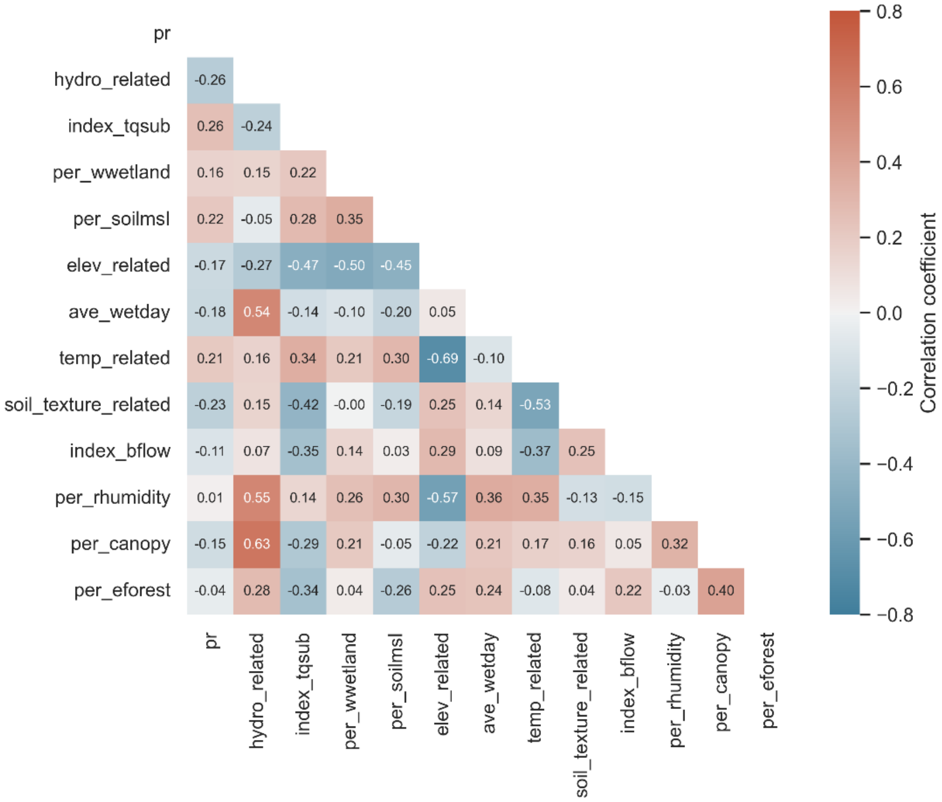

Figure 4. Covariance heatmap of P_r and the 12 selected NHDPlus predictors.

3.3 P_r map

We develop a spatially continuous map of P_r over CONUS by applying the final XGBoost model over the 2.6 million NHDPlus local catchments, as shown in Figure 5. The spatial patterns of P_r are generally consistent with those in Figure 1. High P_r values, shown in orange and red, are mostly located on the southeast coasts, New Mexico, Arizona, southern California, and North Dakota. Low P_r values, shown in blue and purple, are more prevalent in the Northeast and Northwest regions. This consistency between Figures 1 and 5 again confirms that the 2595 independent catchments used in the ML modeling are representative of the whole CONUS domain, hence supporting the transferability of the ML modeling results.

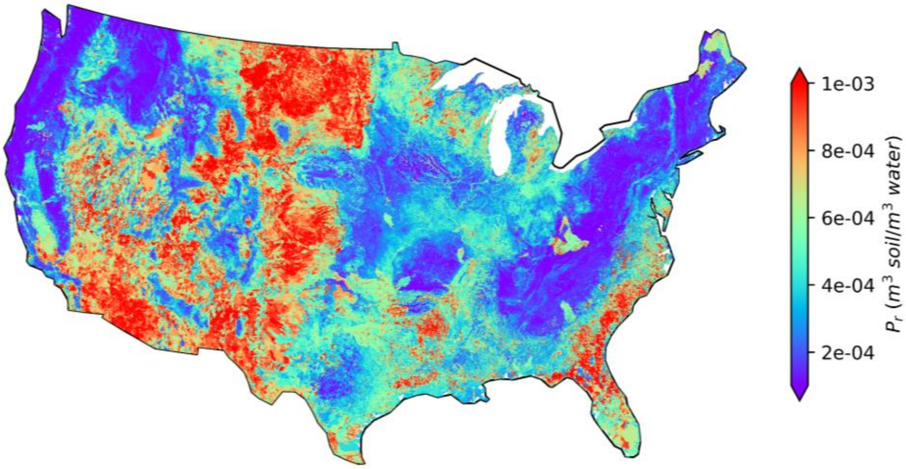

Figure 5. ML model simulated P_r at over 2.6 million NHDPlus local catchments.

3.4 Evaluation

We evaluate the P_r map by comparing the DOC concentration values derived from this map (and Eqn. 6) with those observed, since there is no direct measurement of P_r. The 3210 evaluation gauges (and their corresponding, headwater catchments. See Fig. 1b) are used for this purpose. Note that each of these 3210 evaluation catchments may encompass multiple NHDPlus local catchments. The evaluation thus takes two steps: 1) For each evaluation catchment, calculate its average P_r value by taking the area-weighted average of the local P_r values from the few NHDPlus local catchments located within this catchment; 2) Derive the DOC concentration value for the evaluation catchment (whose outlet is an observational gauge) by using the average P_r value and Eqn. (6); 3) Compare the "derived" DOC concentration with the observed value at the same evaluation catchment. Note that two evaluation catchments are dropped during Step (1) for containing some NHDPlus local catchments without effective model simulated P_r.

Figure 6 shows that our derived DOC concentration values effectively reproduce the spatial variability in the observed values. 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). The MASE, NRMSE and R^2 values are 0.73, 1.81, and 0.47, respectively, further suggesting a satisfactory performance. The scattering only occurs to a small portion of the dots, as indicated by the reddish colours. This scattering may stem from several causes, such as the limited availability of DOC observation data and the uncertainties in model development (see Section 4 for more details). Despite the scattering, the overall alignment between observed and predicted values suggests that our methods, including the generic formula and ML modelling, are appropriate and effective.

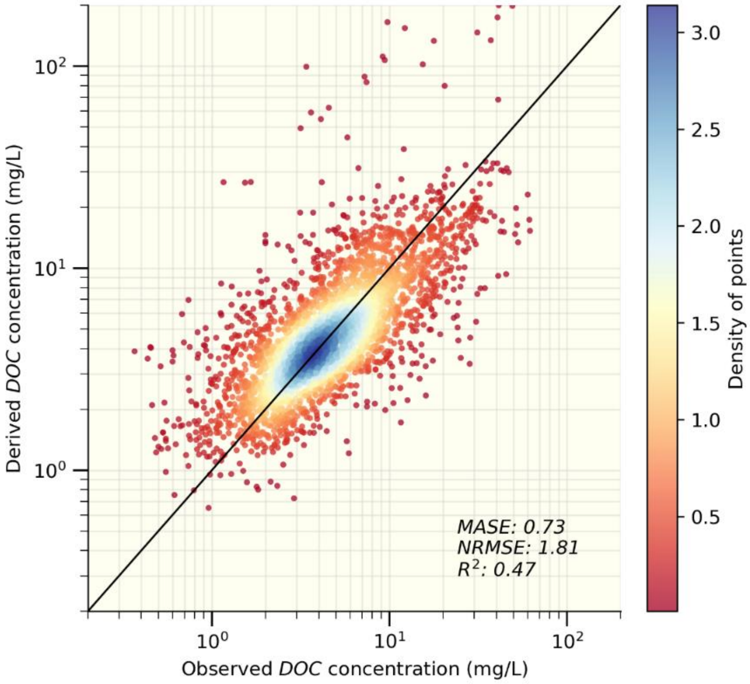

Figure 6. Evaluation of derived DOC concentration at the catchment scale (n=3208). The solid black line indicates a 1:1 ratio. The varying colours indicate the density of points in the scatter plot.

4 Uncertainty analyses

The final product, our P_r map, is subject to uncertainties from various sources. In this study, we have implemented several measures to constrain the uncertainties embedded in the input data and ML modeling exercise. We also look into the ML model parameter uncertainty via sensitivity analyses.

4.1 Efforts to constrain uncertainty

4.1.1 ML model input data

The estimation of the DOC long-term average transformation rate, P_r, relies on SOC data from the HWSD dataset and DOC data from the WQP stations. Despite implementing stringent catchment selection (see Section 2.2.1), the challenge of balancing data quantity and quality persists due to limited DOC measurements. Larger uncertainties in P_r are anticipated in catchments with fewer samples or those where most samples are collected in a single season. Additionally, potential uncertainties in the P_r estimation may arise from the mismatch in sampling periods between SOC and DOC datasets. It is crucial to recognize and account for these uncertainties when interpreting and using the P_r map.

The flowline and catchment attributes from NHDPlus constitute the primary inputs in both training and prediction phases for the ML model, and thus may contribute to the uncertainty in the results. NHDPlus catchment attributes are drawn from diverse sources, including remote sensing data and model simulations. Upstream-accumulated values are derived based on flowline data (Wieczorek et al., 2018). A majority of attributes have been compared to equivalent variables, when available, in the Geospatial Attributes of Gages for Evaluating Streamflow version II (GAGESII) dataset (Falcone et al., 2010). These comparisons have demonstrated reasonably strong alignment. Inherent uncertainties may still arise from inaccurate flowline and catchment delineation, inaccuracies in the source data, the conversion of data formats (e.g., from grid-based to catchment-based), and so on. Furthermore, instances of missing data or attributes with zero-inflated values (e.g., regions highlighted in white in Figure S5b) from the NHDPlus dataset can complicate accurate data interpolation by the ML model. Despite the use of the sparsity-aware technique within the XGBoost algorithm, adept at handling missing or zero-inflated data to a certain extent (Chen and Guestrin, 2016), the presence of such challenges persists. Overcoming these limitations is beyond this study's scope.

4.1.2 ML model development

In contrast to physical-based models with clearly pre-defined structures, machine learning (ML) models endeavor to discern the optimal structure from input data through the training process. Consequently, uncertainty may emerge at any stage of model development, as detailed in Section 2.3. To mitigate model uncertainty, we employ well-established strategies prevalent in diverse applications (Abeshu et al., 2022; Delavar et al., 2019; Li et al., 2022). These encompass techniques such as transformation of input data, training and testing splits, feature selection, hyperparameter tuning, and cross-validation (refer to previous sections for details). These measures aim to constrain the uncertainties inherent in model development processes and fortify the model's predictive capabilities, for example by refining the interpretability of input data, mitigating the risk of overfitting, enhancing generalization performance, and minimizing the introduction of potentially noisy predictors.

In addition to the commonly adopted strategies in using XGBoost and the other ML techniques, we augment the control of model uncertainty through a representativeness check. This check ensures alignment between the distribution of model parameters used during training and those applied in predictions. This additional step serves to enhance the model's transferability from the training catchment to the broader CONUS domain. To gauge the representativeness of our chosen predictors, we conducted a Cumulative Distribution Function (CDF) comparison for each parameter between the observational dataset (derived from 2595 independent catchments) and the entire CONUS dataset (comprising approximately 2.6 million local catchments in NHDPlus). For this comparison, we assess the relative difference in the 5th, 25th, 50th, 75th, and 95th percentiles between the two CDFs. As an illustration, the relative difference for the 5th percentile is computed as the ratio of the difference between the 5th percentile of the available P_r data and that of the entire CONUS data to their average. Table 4 provides a summary of the CDF comparison of the 15 selected predictors (also see supplementary Figure S6). A predictor is deemed representative of the whole CONUS if the average relative difference is less than 0.75. Following Abeshu et al. (2022), the choice of the 0.75 threshold strikes a balance between maintaining data representativeness and avoiding the exclusion of too many predictors. Three predictors, namely "basin_area", "per_hwetland", and "per_shurb", have failed the representativeness check and are consequently excluded. Note that the ML model performance has only slightly changed after reducing the number of predictors from 15 to 12, as shown in the supplementary Figure S7.

Table 4. Representativeness of XGBoost model input predictors over CONUS.

Attributes	Relative difference in percentiles between P_r-*available* and *whole_conus* data					Average
	5th	25th	50th	75th	95th	
basin_area	1.941	1.728	1.669	1.794	1.900	1.806
per_hwetland	0.667	0.667	0.842	1.144	1.529	0.969
per_shurb	0.353	0.624	1.224	1.482	0.889	0.914
per_canopy	1.684	1.090	0.427	0.080	0.078	0.672
per_wwetland	0.769	0.314	0.461	0.621	0.807	0.594
per_eforest	0.667	0.559	0.651	0.502	0.225	0.521
elev_related	0.769	0.806	0.320	0.621	0.008	0.505
hydro_related	0.584	0.898	0.316	0.108	0.106	0.402
per_soilmsI	0.955	0.264	0.152	0.095	0.255	0.344
index_tqsub	0.166	0.135	0.248	0.292	0.393	0.247
index_bflow	0.476	0.304	0.152	0.002	0.027	0.192
per_rhumidity	0.197	0.103	0.015	0.014	0.014	0.068
soil_texture_related	0.095	0.071	0.068	0.071	0.015	0.064
ave_wetday	0.063	0.065	0.028	0.053	0.033	0.048
temp_related	0.035	0.034	0.009	0.029	0.006	0.023

4.2 Sensitivity analyses

Model sensitivity analysis (SA) involves probing the importance of uncertainties in model parameters (Loucks and Van Beek, 2017). We examine our model's sensitivity to each selected predictor using two different methods: 1) dropping one predictor at a time and tracking the changes in model performance, and 2) the Sobol sensitivity analysis approach (Sobol, 2001). Figure 7 demonstrates the model performance difference in the training and testing phases after dropping one of the 12 variables. 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. A 5% threshold is chosen to determine the significance of the change. In general, the shifting pattern in MASE scores remains consistent between the training and testing phases. However, the alterations in MASE values for most predictors, particularly during the testing phase, are minimal or even negligible. In other words, the model appears to be insensitive to most predictors according to this first sensitivity analysis method.

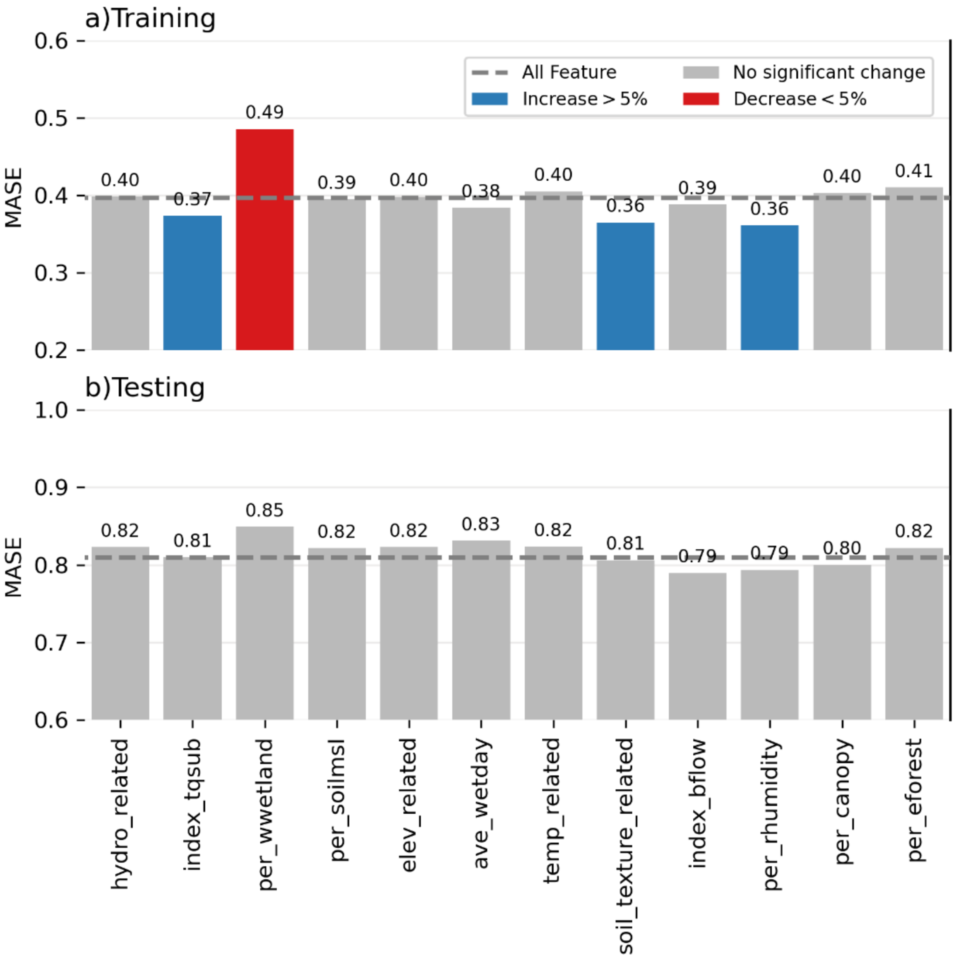

Figure 7. Sensitivity of XGBoost model to predictors in the training and testing phases. The MASE value is represented by the blue, red, and grey bars, indicating whether the model performance increases, decreases, or remains relatively unchanged after dropping the corresponding predictor. The dashed grey line indicates the model performance with all variables included.

The Sobol sensitivity analysis is a widely used variance-based global sensitivity analysis method (Borgonovo and Plischke, 2016). It provides two indices: First-order Index (S1), which measures the sensitivity of an individual predictor itself (local variance), and Total Index (ST), which accounts for the effects of both an individual predictor itself and its interactions with any other predictors (global variance) (Saltelli, 2002; Saltelli et al., 2010). These interactions, which can be of any order, can be isolated. For instance, second and higher-order interactions can be isolated by subtracting SI from ST. The results from the Sobol test are summarized in Table 5. The distribution of S1 is highly right-skewed, suggesting that the model exhibits insensitivity to most predictors if only local variance is considered. There are, however, a few exceptions, such as "hydro_related", and "temp_related", which present high S1 values. The global variance, represented by the ST index, paints a somewhat different picture. When considering the ST index, a broad set of predictors emerge as sensitive, particularly those with ST values exceeding 0.1. It's worth noting that these predictors also hold high rankings in the predictor selection, as shown in Table 2. Furthermore, it is significant that 11 out of the total 12 predictors show a normalized difference between S1 and ST (calculated as (ST-S1)/ST) greater than 50%. This observation underscores the significant interactions among the predictors (Saltelli et al., 2010). This suggests that if a predictor is dropped, the remaining predictors could potentially compensate for its absence, highlighting the nonlinear, high-order interdependence among the predictors in our model.

Table 5. Sobol sensitivity analysis results for the 12 selected predictors.

Predictors	Total Indices (ST)	First Order Indices (S1)	Difference ((ST-S1)/ST)
hydro_related	0.466	0.291	0.375
temp_related	0.311	0.141	0.546
ave_wetday	0.207	0.044	0.788
index_tqsub	0.143	0.003	0.977
per_canopy	0.132	0.028	0.787
per_wwetland	0.125	0.049	0.608
elev_related	0.087	0.017	0.806
index_bflow	0.072	0.012	0.831
per_rhumidity	0.062	0.010	0.836
soil_texture_related	0.034	0.000	1.000
per_eforest	0.024	0.005	0.798
per_soilmsI	0.013	0.002	0.873

The above sensitivity analyses suggest that our model exhibits low sensitivity to most predictors when considering their individual (local) impact. However, the Sobol sensitivity analysis uncovers a heightened degree of sensitivity in the context of global effects, particularly given the significant interactions among the predictors.

5 Potential use and limitations

The P_r map has several promising uses. For instance, one of the pivotal applications of the P_r map is to estimate the lateral leaching of DOC. Figure 8, as an illustration, shows a C_{DOC_runoff} map over CONUS depicting the long-term average concentration of DOC in the leaching flux at over two million NHDPlus local catchments. This map is derived based on Eqn. (4), leveraging the P_r map in Fig. 6 and the top-layer SOC data from HWDS1.2. 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.

https://doi.org/10.5194/essd-2024-43
Preprint. Discussion started: 2 April 2024

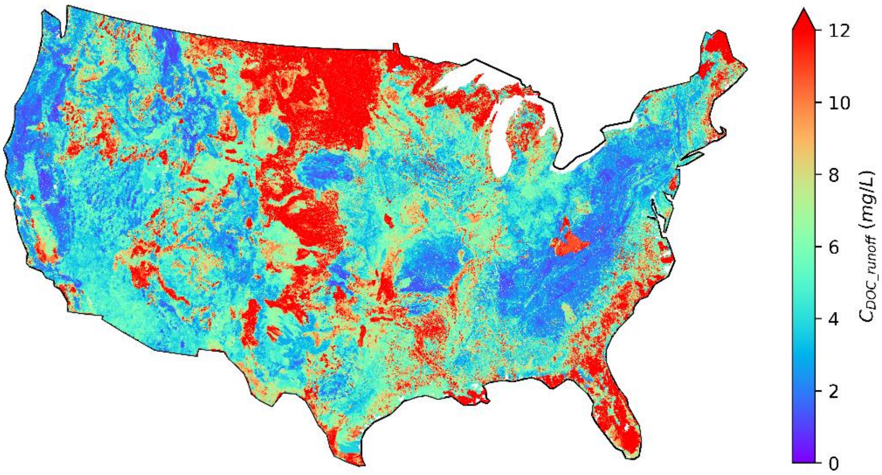

Figure 8. Calculated CONUS map of DOC concentration in leaching flux from soils to headwater streams

The spatial patterns of the C_{DOC_runoff} map are highly correlated to those of the P_r (see Figure 5) and SOC map (see supplementary Figure S5a). Notably, the C_{DOC_runoff} values are high in regions with extremely high SOC values. Additionally, the C_{DOC_runoff} values are high in North Dakota, Montana, and southern coasts, where the P_r values are high. Interestingly, the influences of P_r and SOC can counterbalance each other in some places. For instance, in the upper Rocky Mountains, the SOC storage is abundant due to the presence of forests. However, the low temperature in this region hinders microbial activities, resulting in extremely low P_r value. As a result, the concentration of DOC leaching flux is relatively low. Moreover, the spatial coverage of wetlands also appears to be relevant (see supplementary Figure S5b), which is consistent with the suggested crucial role of wetlands in riverine DOC dynamics (Duan et al., 2017; Leibowitz et al., 2023). For instance, high C_{DOC_runoff} values are observed in upper Minnesota, Florida, and Louisiana, where wetlands are prevalent. In places with few wetlands, like Nevada, Arizona, and New Mexico, the leaching flux concentration is considerably lower.

There are at least two other potential uses of the P_r map: 1) It can support large-scale DOC modeling over CONUS or a major river basin. For instance, testing the use of the map within the framework of the Energy Exascale Earth System Model (Golaz et al., 2019; Caldwell et al., 2019; Burrows et al., 2020) is ongoing and will be reported in the near future. 2) It can be used to provide a quick estimation of riverine DOC concentration or flux at any headwater catchments where no DOC observations are available.

We caution the potential users of the P_r map with several limitations in the methods invoked. Firstly, the P_r values in the map account for the spatial heterogeneity of various DOC-related processes and factors only in a long-term average sense owing to the limited data availability, i.e., the SOC reanalysis data are long-term averages, and the observed riverine DOC data are only available at irregular time intervals. While we believe that such a P_r map is a critical step in effectively capturing the spatial heterogeneity of the relevant processes and environmental factors, incorporating their temporal dynamics is beyond the scope of this study and left for future work. Secondly, the ML techniques are not process-based and thus do not yet offer rich insight into the relevant mechanisms. To improve our understanding of the DOC-related processes, the P_r map should be used in conjunction with other observational data, process-based models, and carefully designed numerical experiments. Last but not least, the ML model has been trained with the data in the CONUS domain only, so it may not be transferable beyond CONUS.

6 Data availability

The resulting P_r and C_{DOC_runoff} maps over CONUS are freely available at https://doi.org/10.5281/zenodo.8339372 (Li et al., 2024). The input data are obtained from the water quality portal (https://www.waterqualitydata.us/), NHDPlus (https://www.epa.gov/waterdata/nhdplus-national-data), ScienceBase (https://doi.org/10.5066/F7765D7V) and HWSD v1.2 (https://www.fao.org/soils-portal/data-hub/soil-maps-and-databases/harmonized-world-soil-database-v12/en/).

7 Conclusions

We develop a new map of P_r, the transformation rate from SOC concentration in soil to DOC concentration in the leaching flux, over CONUS. Evaluation of derived DOC concentration at over 3000 headwater stations confirms the robustness of our methodology, including a generic formula linking SOC and DOC via P_r, riverine DOC observations, environmental variables, and the ML techniques that effectively capture high-order, nonlinear relationships between P_r and the environmental variables. Such a map did not exist before and is highly valuable for large-scale DOC modeling and improving our understanding of the DOC-related processes across the land-river continuum.

Author contributions

LL performed the analysis with the inputs from the co-authors, prepared the figures, and wrote the first draft. HL devised the conceptual idea and supervised the study. GA provided frequent assistance in processing the data and developing the model. All the co-authors contributed to the writing.

Competing interests

At least one of the (co-)authors is a member of the editorial board of Earth System Science Data.

Disclaimer

Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Acknowledgments

This research is supported by the Office of Science of the U.S. Department of Energy as part of the Earth System Model Development program area through the Energy Exascale Earth System Model (E3SM) project. The Pacific Northwest National Laboratory is operated by Battelle for the U.S. Department of Energy under Contract DE-AC05-76RL01830.

References

Abeshu, G. W., Li, H.-Y., Zhu, Z., Tan, Z., and Leung, L. R.: Median bed-material sediment particle size across rivers in the contiguous US, Earth Syst Sci Data, 14, 929–942, https://doi.org/10.5194/essd-14-929-2022, 2022.

Afan, H. A., El-shafie, A., Mohtar, W. H. M. W., and Yaseen, Z. M.: Past, present and prospect of an Artificial Intelligence (AI) based model for sediment transport prediction, J Hydrol (Amst), 541, 902–913, https://doi.org/10.1016/j.jhydrol.2016.07.048, 2016.

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M.: Optuna: A Next-generation Hyperparameter Optimization Framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2623–2631, https://doi.org/10.1145/3292500.3330701, 2019.

Alebachew, M. A., Ye, S., Li, H., Huang, M., Leung, L. R., Fiori, A., and Sivapalan, M.: Regionalization of subsurface stormflow parameters of hydrologic models: Up-scaling from physically based numerical simulations at hillslope scale, J Hydrol (Amst), 519, 683–698, https://doi.org/10.1016/j.jhydrol.2014.07.018, 2014.

Autio, I., Soinne, H., Helin, J., Asmala, E., and Hoikkala, L.: Effect of catchment land use and soil type on the concentration, quality, and bacterial degradation of riverine dissolved organic matter, Ambio, 45, 331–349, https://doi.org/10.1007/s13280-015-0724-y, 2016.

Ayata, S.-D., Irisson, J.-O., Aubert, A., Berline, L., Dutay, J.-C., Mayot, N., Nieblas, A.-E., D'Ortenzio, F., Palmiéri, J., Reygondeau, G., Rossi, V., and Guieu, C.: Regionalisation of the Mediterranean basin, a MERMEX synthesis, Prog Oceanogr, 163, 7–20, https://doi.org/10.1016/j.pocean.2017.09.016, 2018.

Baum, A., Rixen, T., and Samiaji, J.: Relevance of peat draining rivers in central Sumatra for the riverine input of dissolved organic carbon into the ocean, Estuar Coast Shelf Sci, 73, 563–570, https://doi.org/10.1016/j.ecss.2007.02.012, 2007.

Borgonovo, E. and Plischke, E.: Sensitivity analysis: A review of recent advances, Eur J Oper Res, 248, 869–887, https://doi.org/10.1016/j.ejor.2015.06.032, 2016.

Brooks, P. D., McKnight, D. M., and Bencala, K. E.: The relationship between soil heterotrophic activity, soil dissolved organic carbon (DOC) leachate, and catchment-scale DOC export in headwater catchments, Water Resour Res, 35, 1895–1902, https://doi.org/10.1029/1998WR900125, 1999.

Burrows, S. M., Maltrud, M., Yang, X., Zhu, Q., Jeffery, N., Shi, X., Ricciuto, D., Wang, S., Bisht, G., Tang, J., Wolfe, J., Harrop, B. E., Singh, B., Brent, L., Baldwin, S., Zhou, T., Cameron-Smith, P., Keen, N., Collier, N., Xu, M., Hunke, E. C., Elliott, S. M., Turner, A. K., Li, H., Wang, H., Golaz, J. -C., Bond-Lamberty, B., Hoffman, F. M., Riley, W. J., Thornton, P. E., Calvin, K., and Leung, L. R.: The DOE E3SM v1.1 Biogeochemistry Configuration: Description and Simulated Ecosystem-Climate Responses to Historical Changes in Forcing, J Adv Model Earth Syst, 12, https://doi.org/10.1029/2019MS001766, 2020.

Caldwell, P. M., Mametjanov, A., Tang, Q., Van Roekel, L. P., Golaz, J. C., Lin, W., Bader, D. C., Keen, N. D., Feng, Y., Jacob, R., Maltrud, M. E., Roberts, A. F., Taylor, M. A., Veneziani, M., Wang, H., Wolfe, J. D., Balaguru, K., Cameron-Smith, P., Dong, L., Klein, S. A., Leung, L. R., Li, H. Y., Li, Q., Liu, X., Neale, R. B., Pinheiro, M., Qian, Y., Ullrich, P. A., Xie, S., Yang, Y., Zhang, Y., Zhang, K., and Zhou, T.: The DOE E3SM Coupled Model Version 1: Description and Results at High Resolution, J Adv Model Earth Syst, 11, 4095–4146, https://doi.org/10.1029/2019MS001870, 2019.

Camino-Serrano, M., Gielen, B., Luyssaert, S., Ciais, P., Vicca, S., Guenet, B., Vos, B. De, 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., Nilsson, M. B., Peichl, M., Schwendenmann, L., Siemens, J., and Janssens, I.: Linking variability in soil solution dissolved organic carbon to climate, soil type, and vegetation type, Global Biogeochem Cycles, 28, 497–509, https://doi.org/10.1002/2013GB004726, 2014.

Chegini, T., Li, H.-Y., and Leung, L.: HyRiver: Hydroclimate Data Retriever, J Open Source Softw, 6, 3175, https://doi.org/10.21105/joss.03175, 2021.

Chen, T. and Guestrin, C.: XGBoost: A scalable tree boosting system, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794, https://doi.org/10.1145/2939672.2939785, 2016.

Chow, V. Te, Maidment, D. R., and Mays, L. W.: Applied hydrology, McGraw-Hill, 572 pp., 1988.

Daoud, J. I.: Multicollinearity and Regression Analysis, in: Journal of Physics: Conference Series, https://doi.org/10.1088/1742-6596/949/1/012009, 2018.

Delavar, M. R., Gholami, A., Shiran, G. R., Rashidi, Y., Nakhaeizadeh, G. R., Fedra, K., and Afshar, S. H.: A novel method for improving air pollution prediction based on machine learning approaches: A case study applied to the capital city of Tehran, ISPRS Int J Geoinf, 8, https://doi.org/10.3390/ijgi8020099, 2019.

Doron, M., Brasseur, P., and Brankart, J. M.: Stochastic estimation of biogeochemical parameters of a 3D ocean coupled physical-biogeochemical model: Twin experiments, Journal of Marine Systems, 87, 194–207, https://doi.org/10.1016/j.jmarsys.2011.04.001, 2011.

Duan, S., He, Y., Kaushal, S. S., Bianchi, T. S., Ward, N. D., and Guo, L.: Impact of Wetland Decline on Decreasing Dissolved Organic Carbon Concentrations along the Mississippi River Continuum, Front Mar Sci, 3, https://doi.org/10.3389/fmars.2016.00280, 2017.

Duarte, C. M.: Reviews and syntheses: Hidden forests, the role of vegetated coastal habitats in the ocean carbon budget, https://doi.org/10.5194/bg-14-301-2017, 23 January 2017.

Ducharne, A., Golaz, C., Leblois, E., Laval, K., Polcher, J., Ledoux, E., and De Marsily, G.: Development of a high resolution runoff routing model, calibration and application to assess runoff from the LMD GCM, J Hydrol (Amst), 280, 207–228, https://doi.org/10.1016/S0022-1694(03)00230-0, 2003.

Dupas, R., Curie, F., Gascuel-Odoux, C., Moatar, F., Delmas, M., Parnaudeau, V., and Durand, P.: Assessing N emissions in surface water at the national level: Comparison of country-wide vs. regionalized models, Science of the Total Environment, 443, 152–162, https://doi.org/10.1016/j.scitotenv.2012.10.011, 2013.

Falcone, J. A., Carlisle, D. M., Wolock, D. M., and Meador, M. R.: GAGES: A stream gage database for evaluating natural and altered flow conditions in the conterminous United States, Data Papers Ecology, 621 pp., 2010.

Fan, C., Song, C., Liu, K., Ke, L., Xue, B., Chen, T., Fu, C., and Cheng, J.: Century-Scale Reconstruction of Water Storage Changes of the Largest Lake in the Inner Mongolia Plateau Using a Machine Learning Approach, Water Resour Res, 57, https://doi.org/10.1029/2020WR028831, 2021.

Finlay, J., Neff, J., Zimov, S., Davydova, A., and Davydov, S.: Snowmelt dominance of dissolved organic carbon in high-latitude watersheds: Implications for characterization and flux of river DOC, Geophys Res Lett, 33, https://doi.org/10.1029/2006GL025754, 2006.

Fischer, G., Nachtergaele, F., Prieler, S., Van Velthuizen, H. T., Verelst, L., and Wiberg, D.: Global Agro-ecological Zones Assessment for Agriculture (GAEZ 2008), Laxenburg, Austria and FAO, Rome, Italy, 2008.

Futter, M. N., Butterfield, D., Cosby, B. J., Dillon, P. J., Wade, A. J., and Whitehead, P. G.: Modeling the mechanisms that control in-stream dissolved organic carbon dynamics in upland and forested catchments, Water Resour Res, 43, https://doi.org/10.1029/2006WR004960, 2007.

Golaz, J. C., Caldwell, P. M., Van Roekel, L. P., Petersen, M. R., Tang, Q., Wolfe, J. D., Abeshu, G., Anantharaj, V., Asay-Davis, X. S., Bader, D. C., Baldwin, S. A., Bisht, G., Bogenschutz, P. A., Branstetter, M., Brunke, M. A., Brus, S. R., Burrows, S. M., Cameron-Smith, P. J., Donahue, A. S., Deakin, M., Easter, R. C., Evans, K. J., Feng, Y., Flanner, M., Foucar, J. G., Fyke, J. G., Griffin, B. M., Hannay, C., Harrop, B. E., Hoffman, M. J., Hunke, E. C., Jacob, R. L., Jacobsen, D. W., Jeffery, N., Jones, P. W., Keen, N. D., Klein, S. A., Larson, V. E., Leung, L. R., Li, H. Y., Lin, W., Lipscomb, W. H., Ma, P. L., Mahajan, S., Maltrud, M. E., Mametjanov, A., McClean, J. L., McCoy, R. B., Neale, R. B., Price, S. F., Qian, Y., Rasch, P. J., Reeves Eyre, J. E. J., Riley, W. J., Ringler, T. D., Roberts, A. F., Roesler, E. L., Salinger, A. G., Shaheen, Z., Shi, X., Singh, B., Tang, J., Taylor, M. A., Thornton, P. E., Turner, A. K., Veneziani, M., Wan, H., Wang, H., Wang, S., Williams, D. N., Wolfram, P. J., Worley, P. H., Xie, S., Yang, Y., Yoon, J. H., Zelinka, M. D., Zender, C. S., Zeng, X., Zhang, C., Zhang, K., Zhang, Y., Zheng, X., Zhou, T., and Zhu, Q.: The DOE E3SM Coupled Model Version 1: Overview and Evaluation at Standard Resolution, J Adv Model Earth Syst, 11, 2089–2129, https://doi.org/10.1029/2018MS001603, 2019.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J Hydrol (Amst), 377, 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Hamby, D. M.: A Comparison of Sensitivity Analysis Techniques, Health Phys, 68, 195–204, https://doi.org/10.1097/00004032-199502000-00005, 1995.

Hansell, D., Carlson, C., Repeta, D., and Schlitzer, R.: Dissolved Organic Matter in the Ocean: A Controversy Stimulates New Insights, Oceanography, 22, 202–211, https://doi.org/10.5670/oceanog.2009.109, 2009.

Helton, A. M., Wright, M. S., Bernhardt, E. S., Poole, G. C., Cory, R. M., and Stanford, J. A.: Dissolved organic carbon lability increases with water residence time in the alluvial aquifer of a river floodplain ecosystem, J Geophys Res Biogeosci, 120, 693–706, https://doi.org/10.1002/2014JG002832, 2015.

Hyndman, R. J. and Koehler, A. B.: Another look at measures of forecast accuracy, Int J Forecast, 22, 679–688, https://doi.org/10.1016/j.ijforecast.2006.03.001, 2006.

Jing, X., Tian, G., Li, M., and Javeed, S. A.: Research on the spatial and temporal differences of china's provincial carbon emissions and ecological compensation based on land carbon budget accounting, Int J Environ Res Public Health, 18, https://doi.org/10.3390/ijerph182412892, 2021.

Knoben, W. J. M., Freer, J. E., and Woods, R. A.: Technical note: Inherent benchmark or not? Comparing Nash-Sutcliffe and Kling-Gupta efficiency scores, Hydrol Earth Syst Sci, 23, 4323–4331, https://doi.org/10.5194/hess-23-4323-2019, 2019.

Kortelainen, N. M. and Karhu, J. A.: Tracing the decomposition of dissolved organic carbon in artificial groundwater recharge using carbon isotope ratios, Applied Geochemistry, 21, 547–562, https://doi.org/10.1016/j.apgeochem.2006.01.004, 2006.

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

Leibowitz, S. G., Hill, R. A., Creed, I. F., Compton, J. E., Golden, H. E., Weber, M. H., Rains, M. C., Jones, C. E., Lee, E. H., Christensen, J. R., Bellmore, R. A., and Lane, C. R.: National hydrologic connectivity classification links wetlands with stream water quality, Nature Water, 1, 370–380, https://doi.org/10.1038/s44221-023-00057-w, 2023.

Li, H. Y. and Sivapalan, M.: Functional approach to exploring climatic and landscape controls on runoff generation: 2 Timing of runoff storm response, Water Resour Res, 50, 9323–9342, https://doi.org/10.1002/2014WR016308, 2014.

Li, H. Y., Sivapalan, M., Tian, F., and Harman, C.: Functional approach to exploring climatic and landscape controls of runoff generation: 1. Behavioral constraints on runoff volume, Water Resour Res, 50, 9300–9322, https://doi.org/10.1002/2014WR016307, 2014.

Li, H., Wigmosta, M. S., Wu, H., Huang, M., Ke, Y., Coleman, A. M., and Leung, L. R.: A physically based runoff routing model for land surface and earth system models, J Hydrometeorol, 14, 808–828, https://doi.org/10.1175/JHM-D-12-015.1, 2013.

Li, L., Li, H.-Y., Abeshu, G., Tang, J., Leung, L. R., Liao, C., Tan, Z., Tian, H., Thornton, P., & Yang, X.: Deriving a Transformation Rate Map of Dissolved Organic Carbon over the Contiguous U.S., Zenodo [Data set], https://doi.org/10.5281/zenodo.8339372, 2024.

Li, L., Qiao, J., Yu, G., Wang, L., Li, H. Y., Liao, C., and Zhu, Z.: Interpretable tree-based ensemble model for predicting beach water quality, Water Res, 211, https://doi.org/10.1016/j.watres.2022.118078, 2022.

Li, M., Peng, C., Zhou, X., Yang, Y., Guo, Y., Shi, G., and Zhu, Q.: Modeling Global Riverine DOC Flux Dynamics From 1951 to 2015, J Adv Model Earth Syst, 11, 514–530, https://doi.org/10.1029/2018MS001363, 2019.

Liao, C., Zhuang, Q., Leung, L. R., and Guo, L.: Quantifying Dissolved Organic Carbon Dynamics Using a Three-Dimensional Terrestrial Ecosystem Model at High Spatial-Temporal Resolutions, J Adv Model Earth Syst, 11, 4489–4512, https://doi.org/10.1029/2019MS001792, 2019.

Liu, H., Li, Q., Bai, Y., Yang, C., Wang, J., Zhou, Q., Hu, S., Shi, T., Liao, X., and Wu, G.: Improving satellite retrieval of oceanic particulate organic carbon concentrations using machine learning methods, Remote Sens Environ, 256, https://doi.org/10.1016/j.rse.2021.112316, 2021.

Lønborg, C., Carreira, C., Jickells, T., and Álvarez-Salgado, X. A.: Impacts of Global Change on Ocean Dissolved Organic Carbon (DOC) Cycling, https://doi.org/10.3389/fmars.2020.00466, 23 June 2020.

Loucks, D.P. and Van Beek, E.: Water Resource Systems Planning and Management: An Introduction to Methods, Models, and Applications, Springer International Publishing, 624 pp., ISBN 9783319442341, 2017.

Ludwig, W., Probst, J.-L., and Kempe, S.: Predicting the oceanic input of organic carbon by continental erosion, Global Biogeochem Cycles, 10, 23–41, https://doi.org/10.1029/95GB02925, 1996.

Lundberg, S. M. and Lee, S.-I.: A Unified Approach to Interpreting Model Predictions, in: Advances in Neural Information Processing Systems, 2017.

McKay, L.; Bondelid, T.; Dewald, T.; Johnston, J.; Moore, R.; Rea, A., NHDPlus Version 2: User Guide. 2012.

Metherell, A. K., Harding, L. A., Cole, C. V., and Parton, W. J.: CENTURY Soil Organic Matter Model Environment. Technical Documentation Agroecosystem Version 4.0. Great Plains System Research Unit. Technical Report No. 4., Fort Collins, 1993.

Nakhavali, M., Friedlingstein, P., Lauerwald, R., Tang, J., Chadburn, S., Camino-Serrano, M., Guenet, B., Harper, A., Walmsley, D., Peichl, M., and Gielen, B.: Representation of dissolved organic carbon in the JULES land surface model (vn4.4-JULES-DOCM), Geosci Model Dev, 11, 593–609, https://doi.org/10.5194/gmd-11-593-2018, 2018.

Parton, W. J., Hartman, M., Ojima, D., and Schimel, D.: DAYCENT and its land surface submodel: description and testing, Global and Planetary Change, 35–48 pp., 1998.

Parton, W. J., Schimel, D. S., Cole, C. V., and Ojima, D. S.: Analysis of Factors Controlling Soil Organic Matter Levels in Great Plains Grasslands, Soil Science Society of America Journal, 51, 1173–1179, https://doi.org/10.2136/sssaj1987.03615995005100050015x, 1987.

Pool, S., Vis, M., and Seibert, J.: Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency, Hydrological Sciences Journal, 63, 1941–1953, https://doi.org/10.1080/02626667.2018.1552002, 2018.

Romeiko, X. X., Guo, Z., Pang, Y., Lee, E. K., and Zhang, X.: Comparing machine learning approaches for predicting spatially explicit life cycle global warming and eutrophication impacts from corn production, Sustainability (Switzerland), 12, https://doi.org/10.3390/su12041481, 2020.

Ross, C. W., Prihodko, L., Anchang, J., Kumar, S., Ji, W., and Hanan, N. P.: HYSOGs250m, global gridded hydrologic soil groups for curve-number-based runoff modeling, Sci Data, 5, 180091, https://doi.org/10.1038/sdata.2018.91, 2018.

Saltelli, A., Annoni, P., Azzini, I., Campolongo, F., Ratto, M., and Tarantola, S.: Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index, Comput Phys Commun, 181, 259–270, https://doi.org/10.1016/j.cpc.2009.09.018, 2010.

Saltelli, A.: Making best use of model evaluations to compute sensitivity indices, Comput Phys Commun, 145, 280–297, https://doi.org/10.1016/S0010-4655(02)00280-1, 2002.

Santos, L., Thirel, G., and Perrin, C.: Technical note: Pitfalls in using log-transformed flows within the KGE criterion, Hydrol Earth Syst Sci, 22, 4583–4591, https://doi.org/10.5194/hess-22-4583-2018, 2018.

Schelker, J., Eklöf, K., Bishop, K., and Laudon, H.: Effects of forestry operations on dissolved organic carbon concentrations and export in boreal first-order streams, J Geophys Res Biogeosci, 117, https://doi.org/10.1029/2011JG001827, 2012.

Sivapalan, M.: Pattern, Process and Function: Elements of a Unified Theory of Hydrology at the Catchment Scale, in: Encyclopedia of Hydrological Sciences, Wiley, https://doi.org/10.1002/0470848944.hsa012, 2005.

Sobol, I. M.: Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates, Mathematics and Computers in Simulation, 271–280 pp., 2001.

Steiner, J. L., Sadler, E. J., Chen, J.-S., Wilson, G., James, D., Vandenberg, B., Ross, J., Oster, T., and Cole, K.: Sustaining the Earth's Watersheds-Agricultural Research Data System: Overview of development and challenges, J Soil Water Conserv, 63, 569–576, https://doi.org/10.2489/jswc.63.6.569, 2008.

Strauss, E. A. and Lamberti, G. A.: Effect of dissolved organic carbon quality on microbial decomposition and nitrification rates in stream sediments, Freshw Biol, 47, 65–74, https://doi.org/10.1046/j.1365-2427.2002.00776.x, 2002.

Tan, Z., Leung, L. R., Li, H. Y., and Cohen, S.: Representing Global Soil Erosion and Sediment Flux in Earth System Models, J Adv Model Earth Syst, 14, https://doi.org/10.1029/2021MS002756, 2022.

Teodoru, C. R., Nyoni, F. C., Borges, A. V., Darchambeau, F., Nyambe, I., and Bouillon, S.: Dynamics of greenhouse gases (CO2, CH4, N2O) along the Zambezi River and major tributaries, and their importance in the riverine carbon budget, Biogeosciences, 12, 2431–2453, https://doi.org/10.5194/bg-12-2431-2015, 2015.

Tian, H., Ren, W., Yang, J., Tao, B., Cai, W. J., Lohrenz, S. E., Hopkinson, C. S., Liu, M., Yang, Q., Lu, C., Zhang, B., Banger, K., Pan, S., He, R., and Xue, Z.: Climate extremes dominating seasonal and interannual variations in carbon export from the Mississippi River Basin, Global Biogeochem Cycles, 29, 1333–1347, https://doi.org/10.1002/2014GB005068, 2015b.

Tian, H., Yang, Q., Najjar, R. G., Ren, W., Friedrichs, M. A. M., Hopkinson, C. S., and Pan, S.: Anthropogenic and climatic influences on carbon fluxes from eastern North America to the Atlantic Ocean: A process-based modeling study, J Geophys Res Biogeosci, 120, 752–772, https://doi.org/10.1002/2014JG002760, 2015a.

Tranvik, L. J. and Jansson, M.: Terrestrial export of organic carbon, Nature, 415, 861–862, https://doi.org/10.1038/415861b, 2002.

U.S. Geological Survey, National Water Information System data available on the World Wide Web (Water-Quality Data for the Nation): https://waterdata.usgs.gov/nwis/qw, last access: 27 January 2024.

USEPA, STOrage and RETrieval Data Available on the World Wide Web (EPA STORET): https://www.epa.gov/waterdata/storage-and-retrieval-and-water-quality-exchange, last access: 27 January 2024.

Verrelst, J., Rivera, J. P., van der Tol, C., Magnani, F., Mohammed, G., and Moreno, J.: Global sensitivity analysis of the SCOPE model: What drives simulated canopy-leaving sun-induced fluorescence?, Remote Sens Environ, 166, 8–21, https://doi.org/10.1016/j.rse.2015.06.002, 2015.

Water Quality Portal. Washington (DC): National Water Quality Monitoring Council, United States Geological Survey (USGS), Environmental Protection Agency (EPA); 2021. https://doi.org/10.5066/P9QRKUVJ.

Wickland, K. P., Aiken, G. R., Butler, K., Dornblaser, M. M., Spencer, R. G. M., and Striegl, R. G.: Biodegradability of dissolved organic carbon in the Yukon River and its tributaries: Seasonality and importance of inorganic nitrogen, Global Biogeochem Cycles, 26, https://doi.org/10.1029/2012GB004342, 2012.

Wieczorek, M. E., Jackson, S. E., and Schwarz, G. E.: Select Attributes for NHDPlus Version 2.1 Reach Catchments and Modified Network Routed Upstream Watersheds for the Conterminous United States (ver. 3.0, January 2021): US Geological Survey data release, https://doi.org/10.5066/F7765D7V, 2018.

Wilson, H. F., Saiers, J. E., Raymond, P. A., and Sobczak, W. V.: Hydrologic Drivers and Seasonality of Dissolved Organic Carbon Concentration, Nitrogen Content, Bioavailability, and Export in a Forested New England Stream, Ecosystems, 16, 604–616, https://doi.org/10.1007/s10021-013-9635-6, 2013.

Yao, Y., Tian, H., Pan, S., Najjar, R. G., Friedrichs, M. A. M., Bian, Z., Li, H. Y., and Hofmann, E. E.: Riverine Carbon Cycling Over the Past Century in the Mid-Atlantic Region of the United States, J Geophys Res Biogeosci, 126, https://doi.org/10.1029/2020JG005968, 2021.

Ye, S., Li, H. Y., Huang, M., Alebachew, M. A., Leng, G., Leung, L. R., Wang, S. wen, and Sivapalan, M.: Regionalization of subsurface stormflow parameters of hydrologic models: Derivation from regional analysis of streamflow recession curves, J Hydrol (Amst), 519, 670–682, https://doi.org/10.1016/j.jhydrol.2014.07.017, 2014.

Yeo, I. K. and Johnson, R. A.: A new family of power transformations to improve normality or symmetry, Biometrika, 87, 954–959, https://doi.org/10.1093/biomet/87.4.954, 2000.

Ying, X.: An Overview of Overfitting and its Solutions, in: Journal of Physics: Conference Series, https://doi.org/10.1088/1742-6596/1168/2/022022, 2019.

Zhang, P.: A novel feature selection method based on global sensitivity analysis with application in machine learning-based prediction model, Applied Soft Computing Journal, 85, https://doi.org/10.1016/j.asoc.2019.105859, 2019.