Supporting information for

Combined effects of stream hydrology and land use on basin-scale hyporheic zone denitrification in the Columbia River Basin

Kyongho Son[1], Yilin Fang[1], Jesus D. Gomez-Velez[2,3], Kyuhyun Byun[4], and Xingyuan Chen[1]

[1]Pacific Northwest National Laboratory, Richland, Washington, USA; [2]Department of Civil and Environmental Engineering, Vanderbilt University, Nashville, Tennessee, USA; [3]Climate Change Science Institute & Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA; [4]Department of Environmental Engineering, Incheon National University, Incheon, 22012, South Korea

Contents of this file

Figure S1 to S17

Table S1 to S3

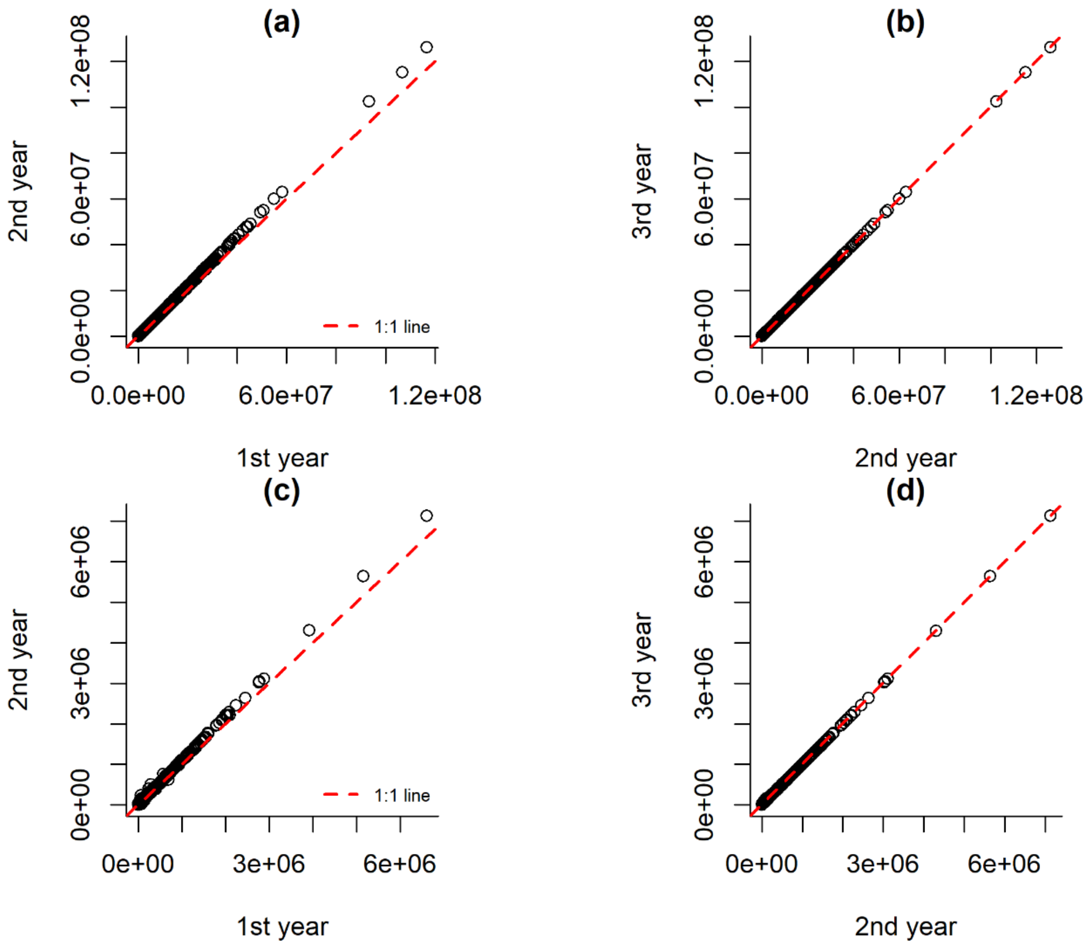

Figure S1. The impact of simulation length on modeled NO_3^- removal amounts (mole N) via vertical and lateral hyporheic exchange: (a) comparison of the 1st year and 2nd year simulation for the vertical modeled NO_3^- removal amounts (mole N); (b) comparison of the 2nd year and 3rd year simulation for the vertical modeled NO_3^- removal amounts; (c) comparison of the 1st year and 2nd year simulation for the lateral modeled NO_3^- removal amounts (mole N); (d) comparison of the 2nd year and 3rd year simulation for the lateral modeled NO_3^- removal amounts.

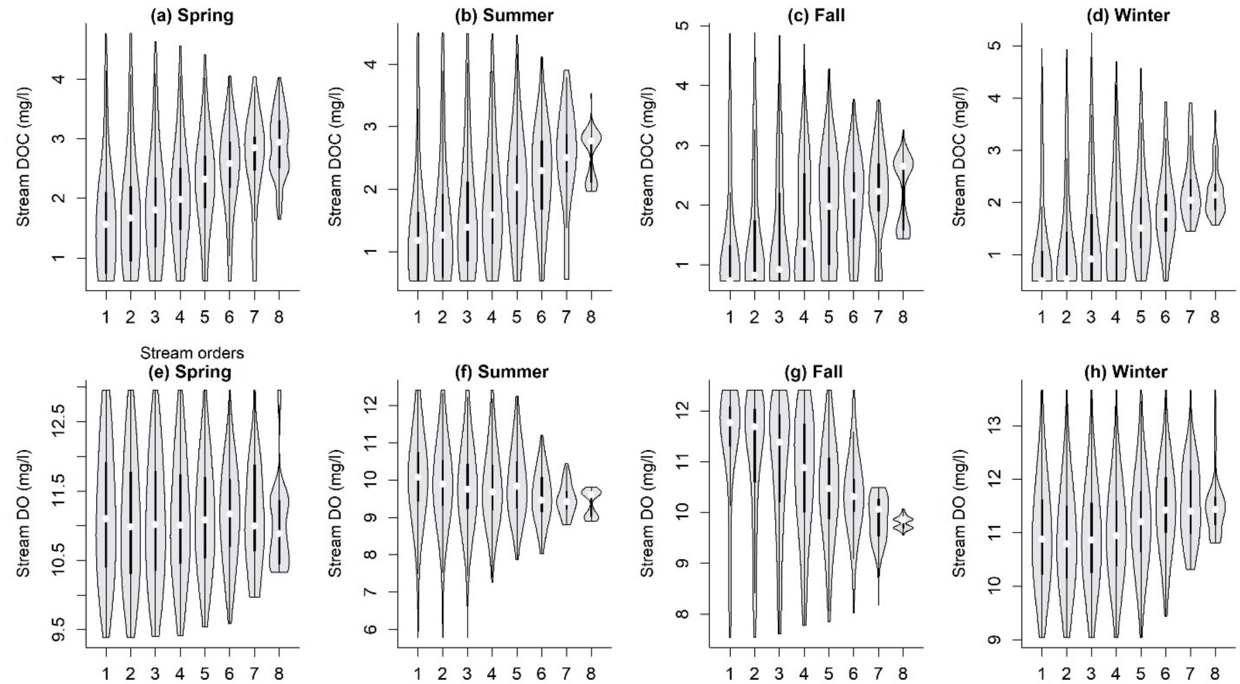

Figure S2. The seasonal stream DOC and DO variations with the stream/river orders.

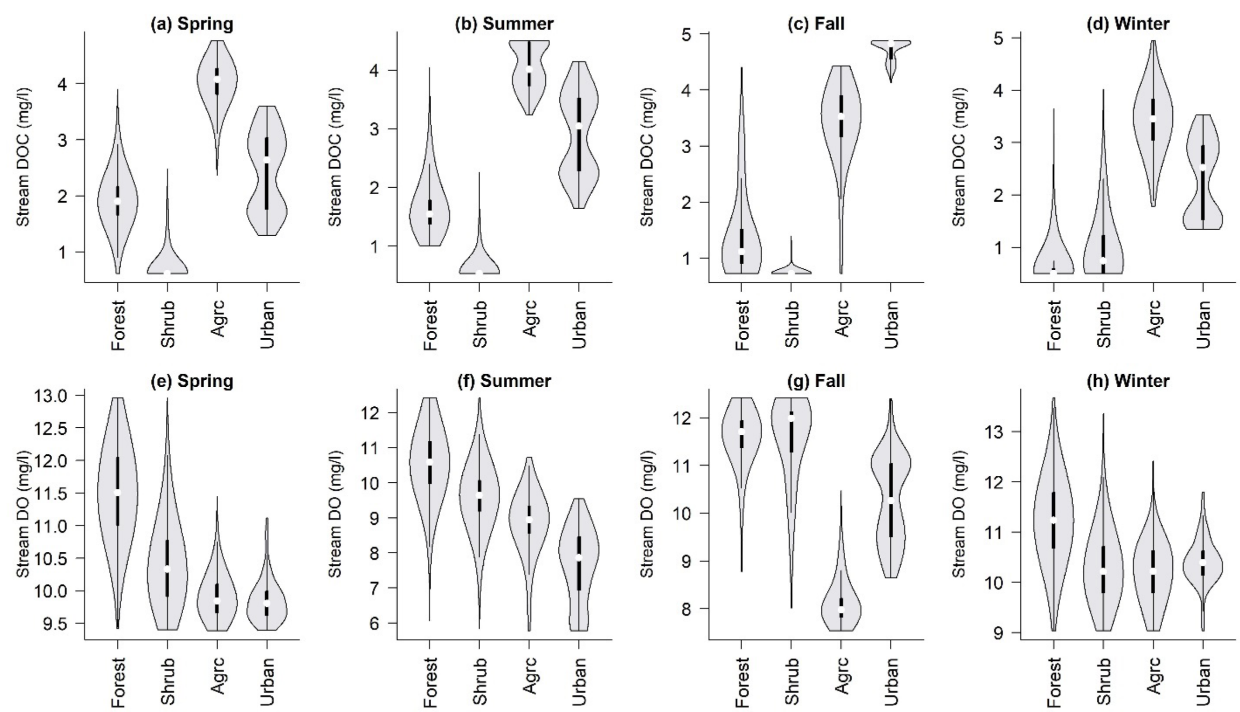

Figure S3. The seasonal stream DOC and DO variations with different land uses.

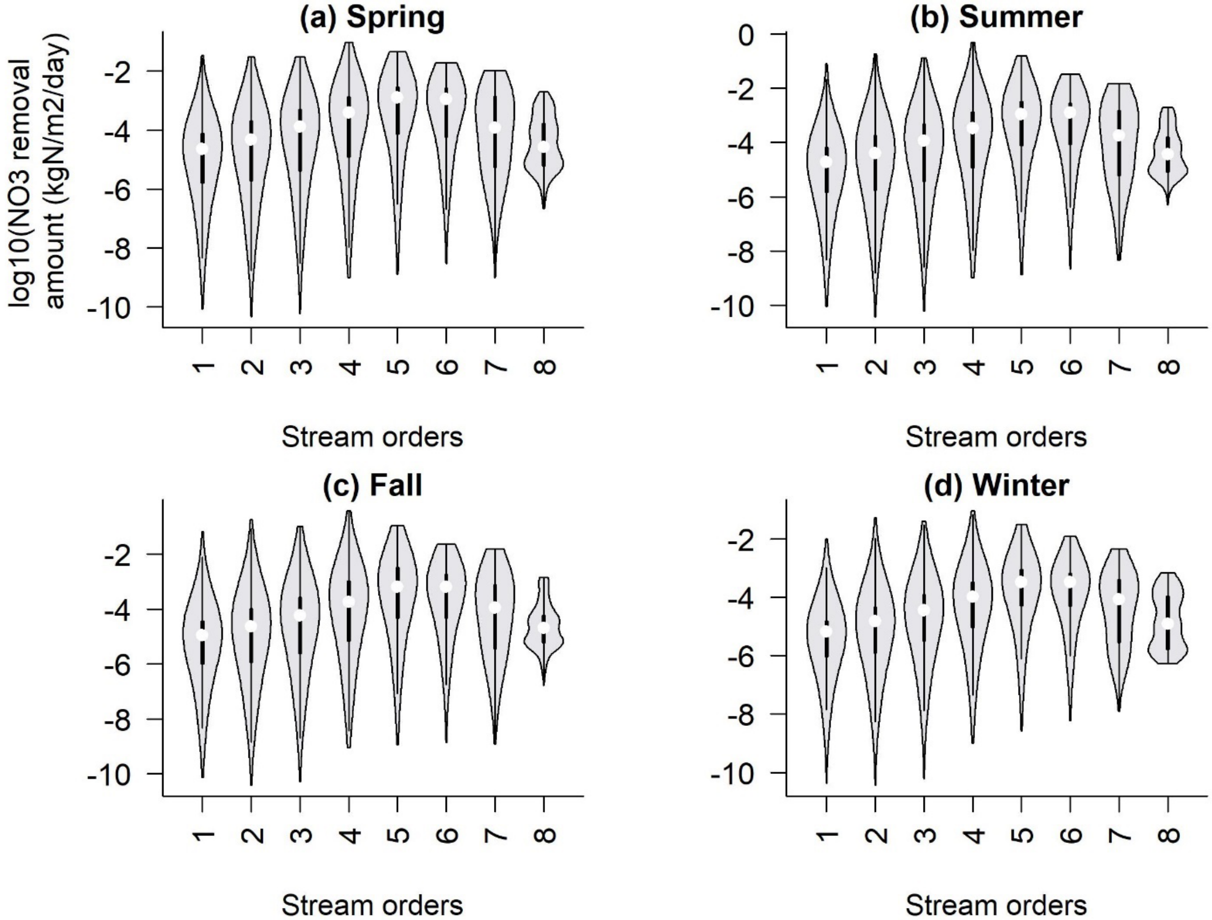

Figure S4. The spatial variation of the modeled HZ NO_3^- removal amounts (kgN/m^2/day) in the reaches with different orders and seasonal substrate concentration inputs.

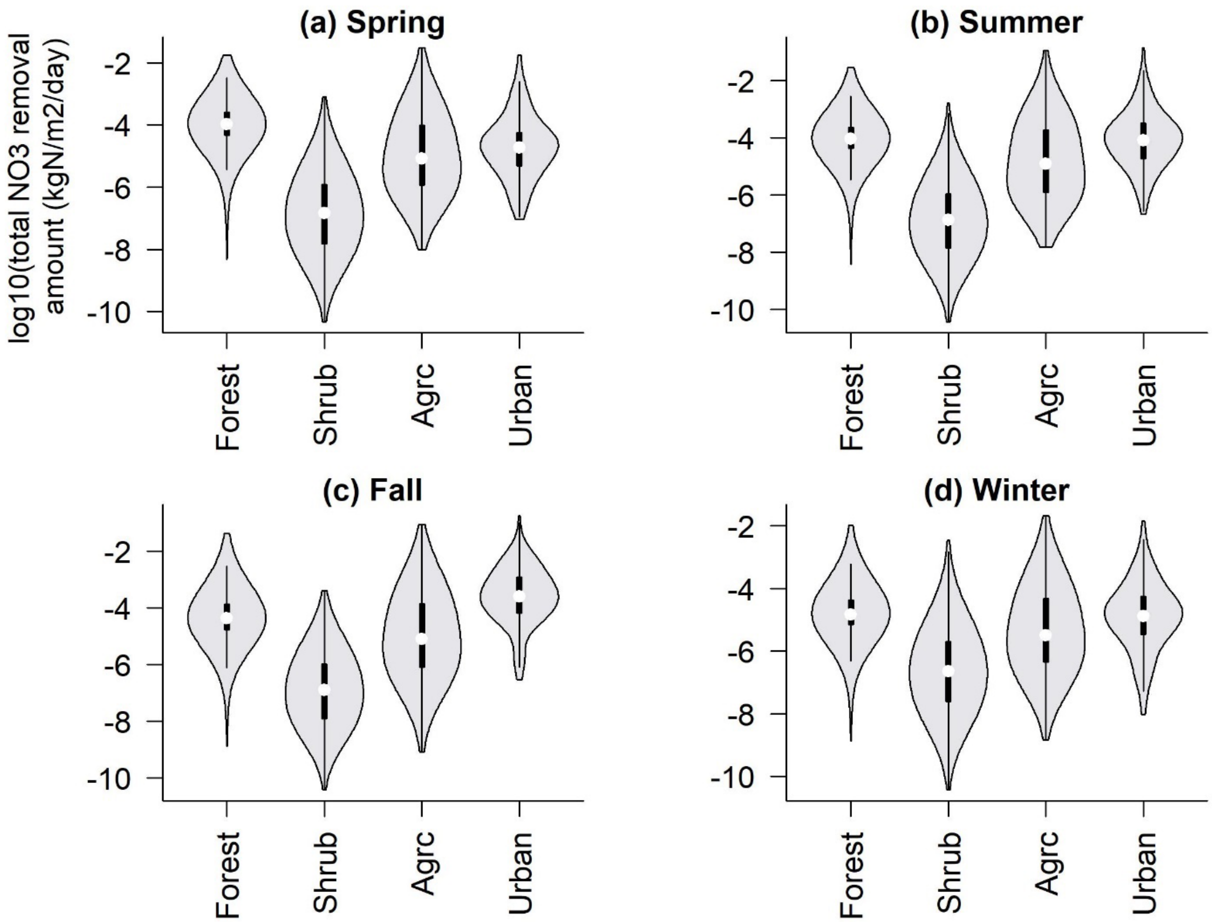

Figure S5. The spatial variation of the modeled HZ NO_3^- removal amounts (kgN/m^2/day) in the reaches with different land uses and seasonal substrate concentration inputs.

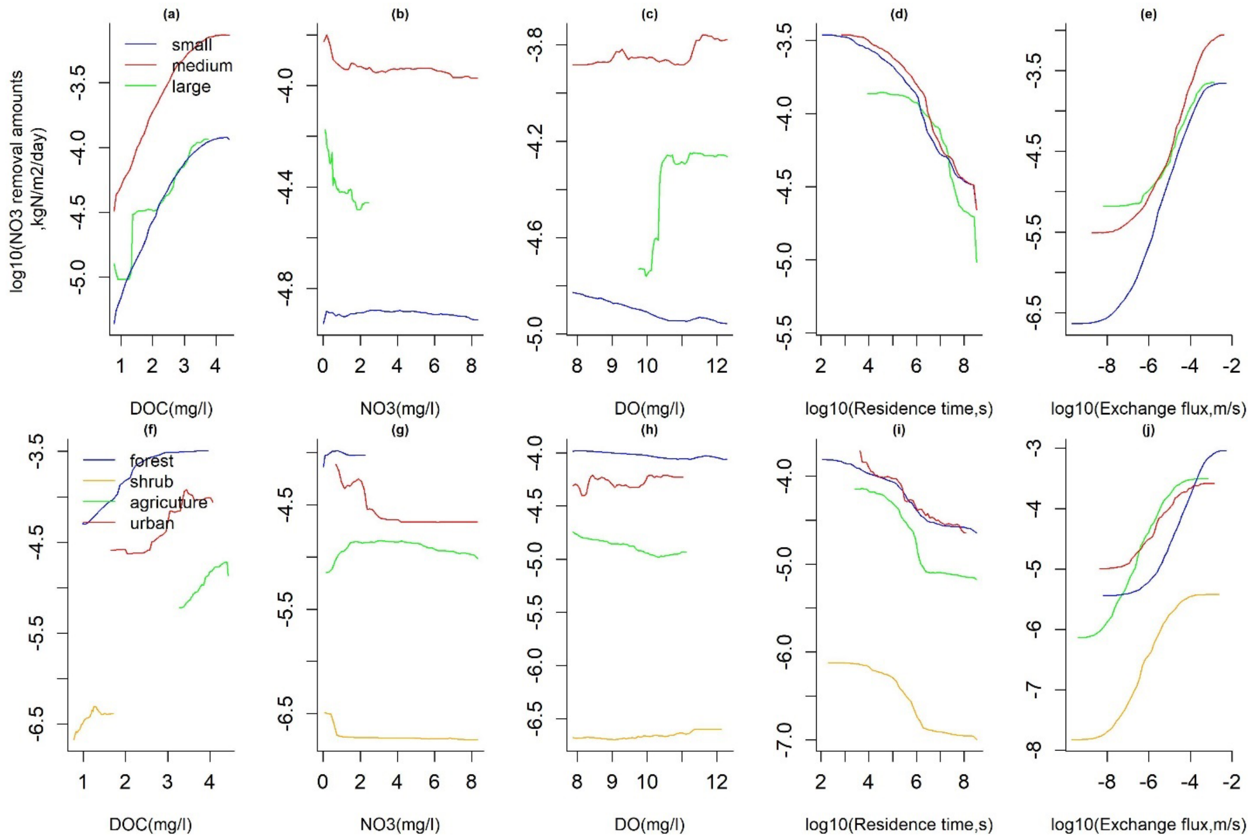

Figure S6. Partial correlation between key model inputs and modeled HZ NO_3^- removal amounts (kgN/m^2/day) in reaches across different sizes and land uses.

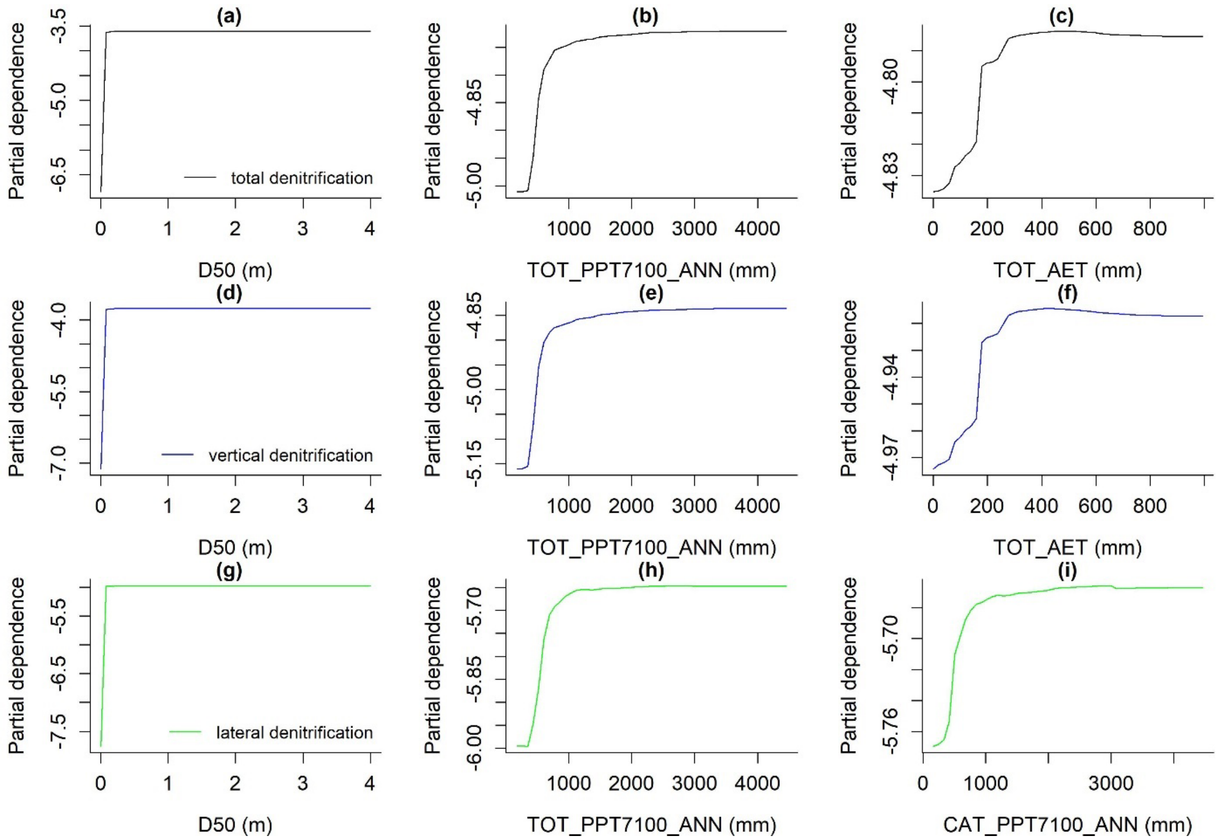

Figure S7. Partial correlation between important variables and modeled NO_3^- removal amounts (kgN/m^2/day): (a, d, g) D50 (median grain size); (b,e,h) TOT_PPT7100_ANN (mean annual precipitation at the NHD cumulative drainage area); (c,f) TOT_AET (mean annual actual evapotranspiration at the NHD cumulative drainage area); and (i) CAT_PPT7100_ANN (mean annual precipitation at the NHD catchment drainage area).

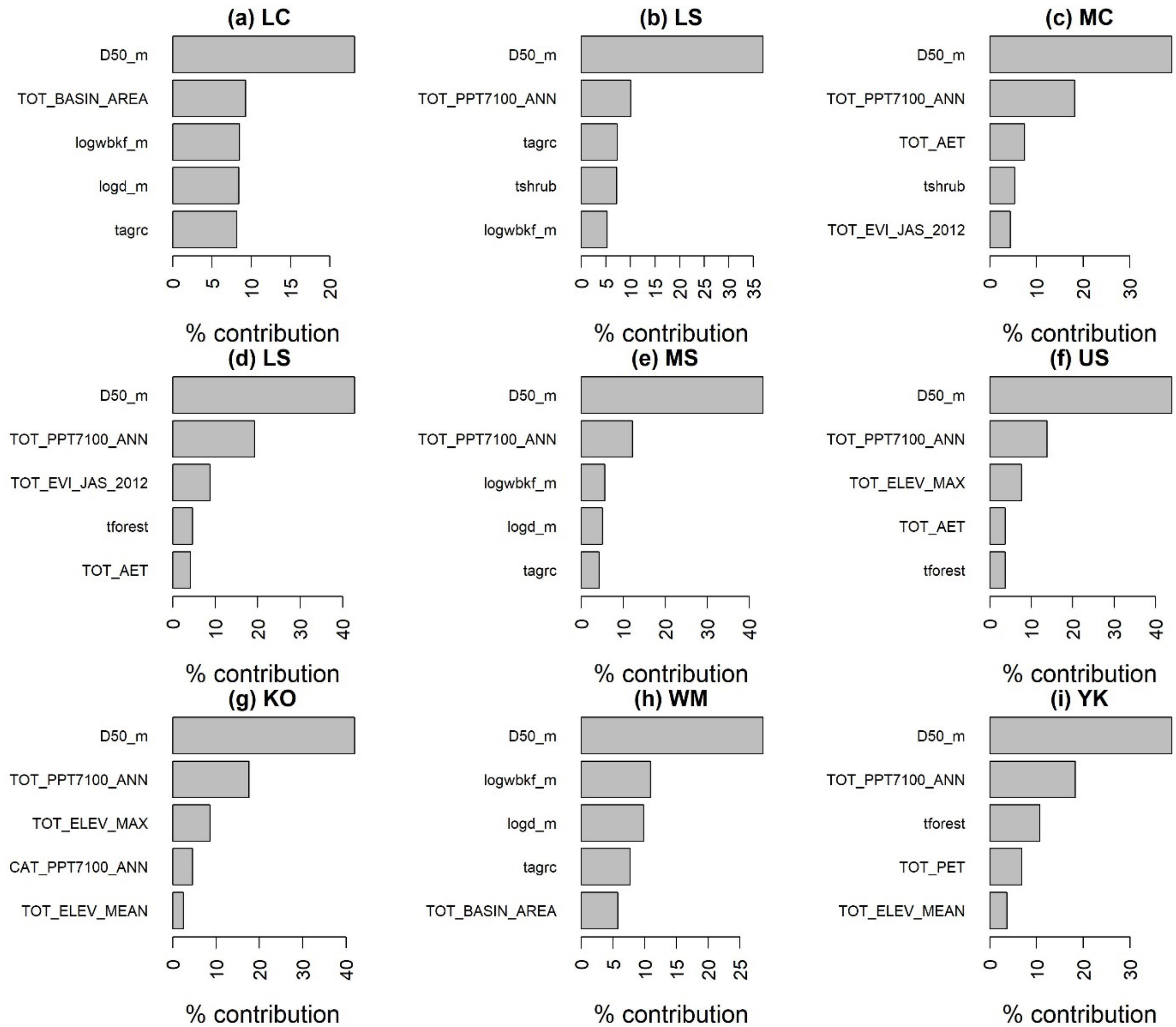

Figure S8. The top five importance variables for total modeled NO_3^- removal amounts (log10, kgN/m^2/day) for the nine sub-basins in the Columbia River Basin: (a) Lower Columbia (LC); (b) Middle Columbia (MC); (c) Upper Columbia (UC); (d) Lower Snake (LS); (e) Middle Snake (MS); (f) Upper Snake (US); (g) Kootenai-Pend Oreille-Spokane (KO); (h) Williamette (WM); and (i) Yakima (YK). D50 is median grain size; TOT_BASIN_AREA is watershed drainage area at the NHD cumulative drainage area; TOT_ELEV_MAX/MEAN is maximum/mean elevation at the NHD cumulative drainage area; logwbkf_m is bankfull width (log10 scale) and logd_m is water depth (log 10 scale); TOT_PET/AET is mean annual potential /actual evapotranspiration at the NHD cumulative drainage area; TOT_PPT7100_ANN is mean annual precipitation at the NHD cumulative drainage area; CAT_PPT7100_ANN is mean annual precipitation at the NHD catchment drainage area; TOT_EVI_JAS_2012 is summer EVI index in year 2012 at the NHD cumulative drainage area; and targrc/tforest/tshrub is the percentage of agricultural/forest/shrub lands at the NHD cumulative drainage area.

Estimating the stream substrate concentrations

Our river corridor model requires stream water DOC, NO_3^-, and DO concentrations at the NHDPLUS reach scale as key substrate concentration inputs. To estimate the stream DOC and DO concentrations, we developed multilinear regression models with the measured stream concentration data, NHDPLUS-based watershed/stream properties (Table S1), and the SPARROW model outputs. For developing the regression model for the stream DOC concentration, we refer to the work of (Yang et al. 2017). The stream DOC concentration data are downloaded from the USGS NWIS (http://waterdata.usgs.gov/nwis) using the "dataretrieve" R package. The lists of gauge stations for the CRB were obtained from the work of (Zarnetske et al., 2018). The period of the samples is from 1/1/1980 to 12/31/2021. The selected stations have both flow and DOC data, their records are longer than 3 years, and least number of samples are 20. The sampled data spanned more than 50% of the observed flow ranges. These conditions help to accurately compute the mean DOC concentration over the various hydrologic conditions. We can find the 65 USGS gauge stations within the CRB, but to use the NHDPLUS watershed/stream reaches database, we only used 55 stations that match with NHDPLUS reach identification number (comid) shown in Figure S9. To predict the annual mean DOC concentration at the NHDPLUS stream reaches of the CRB, we used various watershed properties and variables that may be relevant to the stream DOC concentrations (Table S1). To remove the outlier of the sampled data, we computed the standard deviation (sd) of all sampled data per site, and if the sampled concentration was larger than 3*sd plus mean, the sample was considered an outlier (Yang et al., 2017). Some variables were log-transformed before building the regression model to remove the impact of non-normal variables. For example, soil organic matter (TOT_OM), % wetland (twetland) and dam storage (TOT_NID_STORAGE2010), total nitrogen concentration (tn), annual mean temperature (TOT_TAV7100_ANN), and % clay (TOT_CLAYAVE) were log-transformed. To remove the highly correlated variables, we used a variance inflation factor (VIF) index. If the variable's VIF was larger than 10, we excluded the variable in developing the regression model. Also, when the paired correlation between variables and measured DOC was statistically significant, the variable was included in developing the regression model. The included variables were TOT_SILTAVE, TOT_SANDAVE, CAT_SILTAVE, tshurb, CAT_BFI, logturban, logtargc,logCAT_TAV, and logshurb (Figure S12). We explored the possible combination of multiregression models with the selected variables using the “olsrr” r package (https://cran.r-project.org/web/packages/olsrr/index.html) and found that the regression model using the three variables, tshrub, logtargc, and logshurb, had relatively a high R^2 value (0.469) and a low AIC value (136) compared with other regression models (Figure S11).

Similar to building the annual mean DOC model, we also developed seasonal mean DOC models (Table S2 and Figure S12). The model performance varied with season. The summer DOC model had the lowest model accuracy (R^2=0.359), and the winter DOC model had the highest model accuracy (R^2=0.54). Each model had different variables. The detailed equations of each model are included in Table S2.

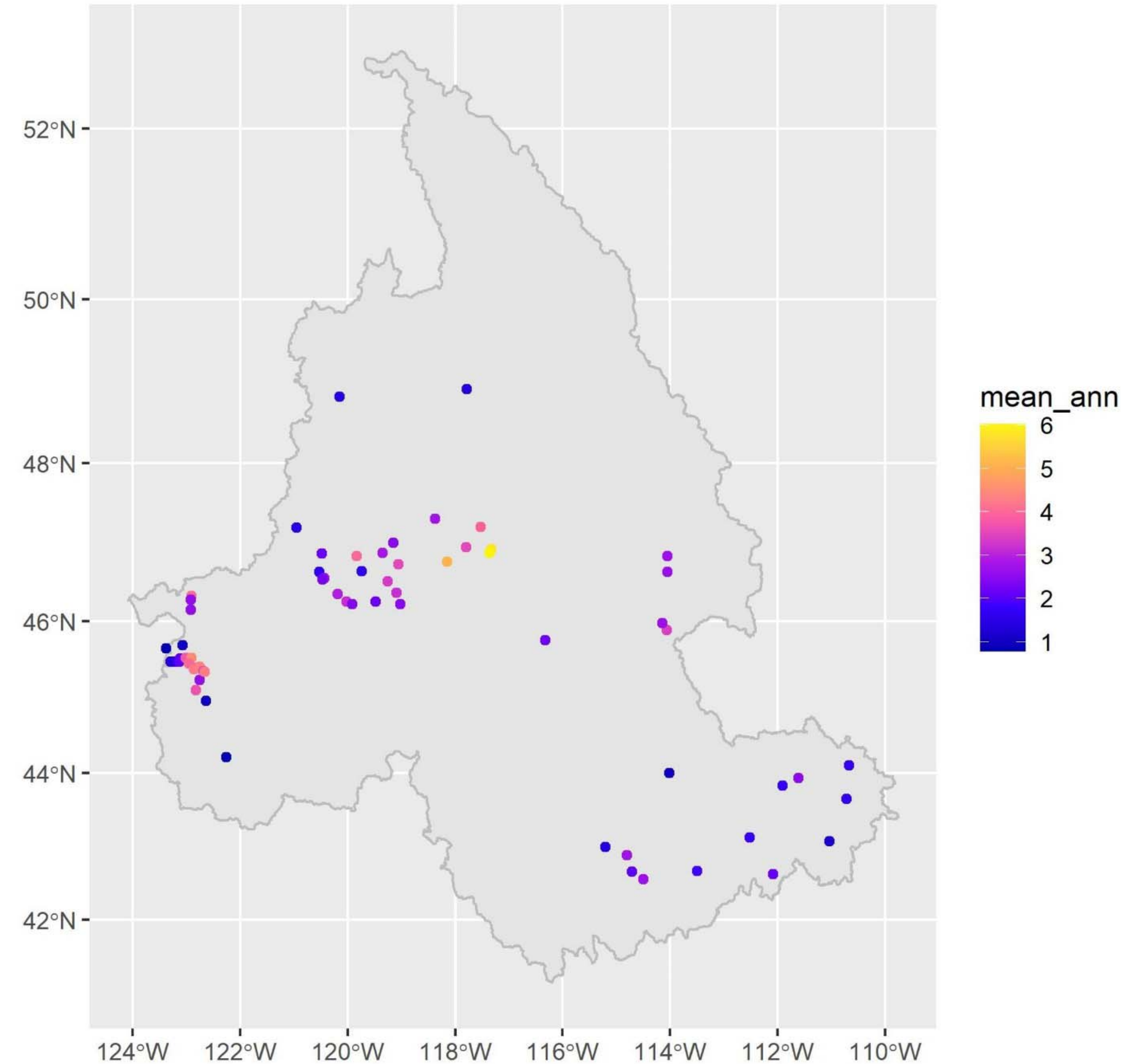

Figure S9. The locations of the used gauge stations and the annual mean stream DOC concentration (mg/l).

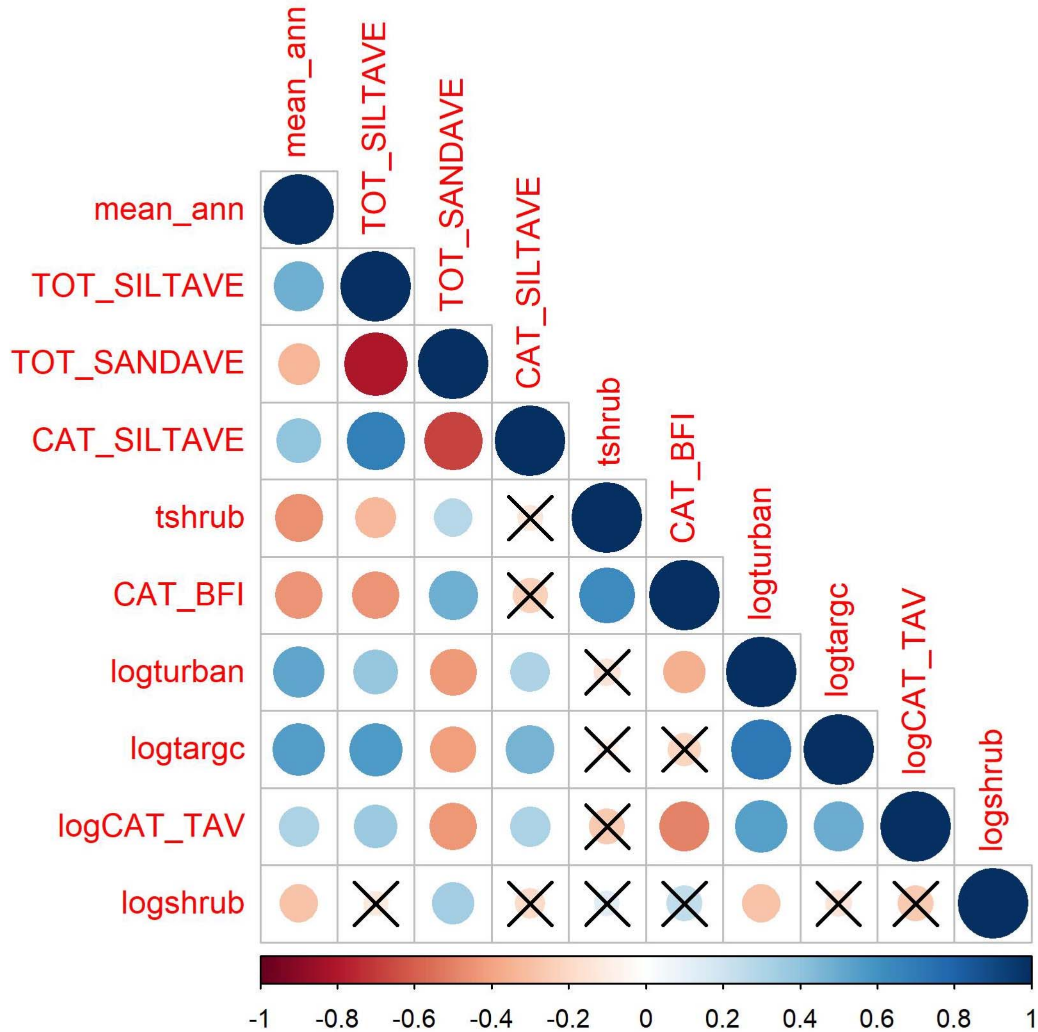

Figure S10. Correlation between selected variables and annual mean DOC concentrations: only variables with the significant (95%) relationship with the annual mean DOC concentration are displayed.

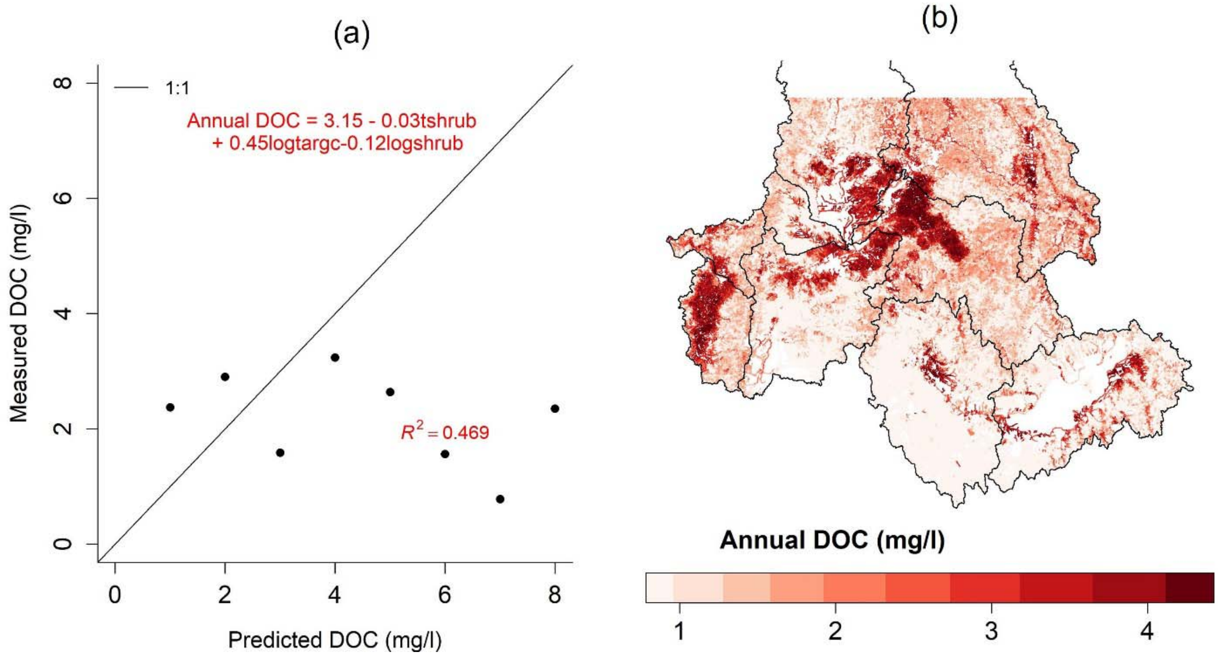

Figure S11. The developed stream annual mean DOC model and its prediction: (a) developed regression model and (b) predicted stream annual mean DOC concentration at the NHDPLUS stream reaches.

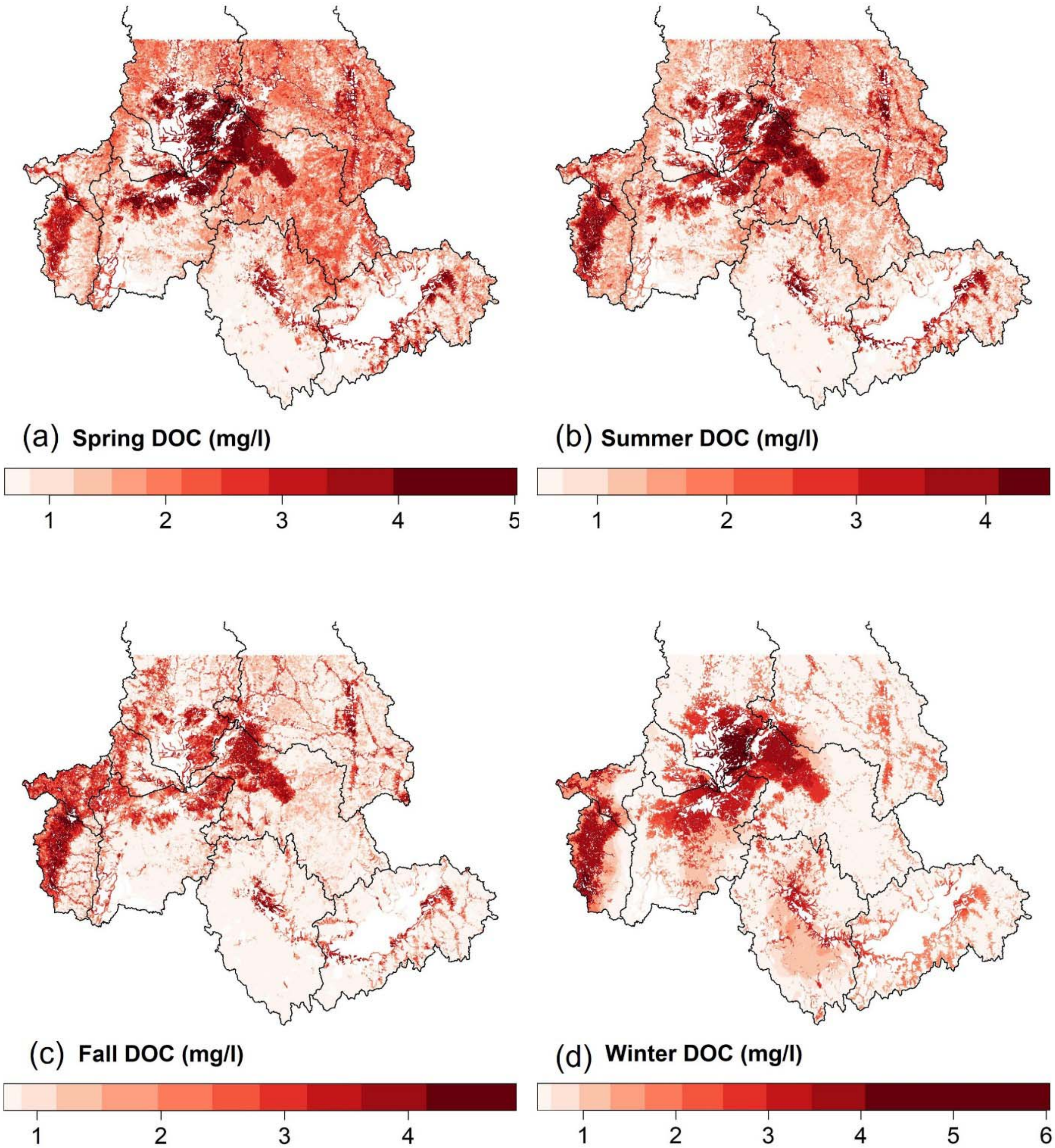

Figure S12. Predicted stream seasonal DOC concentrations at the NHDPLUS stream reaches: (a) spring mean DOC (mg/l); (b) summer mean DOC (mg/l); (c) fall mean DOC (mg/l); and (d) winter mean DOC (mg/l).

To predict stream mean annual DO concentrations at the NHDPLUS stream reaches of the CRB, we used a similar approach to developing the stream DOC regression model. For sampled DO concentration data, the samples collected from 1/1/2007 to 12/31/2021 were downloaded using the “dataretrieve” R package since the DO sensor had some accuracy issues prior to 2007. Another criterion was that the stations should have at least 20 samples to get a reasonable mean concentration over periods. We found 42 gauge stations within the CRB, but only 38 stations matched with the NHDPLUS reach comid. Figure S13 shows the annual mean concentrations of stream DO at the 38 stations in the CRB. A multilinear regression model was developed for predicting stream annual mean DO concentrations at the NHDPLUS stream reaches using various watershed and stream properties and the measured annual mean DO concentration data (Table S1). Figure S14 showed high spatial correlation values between the annual mean DO concentrations and the selected variables. Among the selected variables, tforest, TOT_PPT7100_ANN, logTOT_BASIN_AREA, logTOT_STREAM_SLOPE, and logCAT_NID showed positive correlations with the stream DO concentrations, while TOT_BDAVE, TOT_TWI, logtargc, and logurban showed negative correlations. Also, the selected variables all had low VIF values (<10). We explored the possible combination of multiregression models with the selected variables using the “olsrr” r package. We chose four variables (TOT_BDAVE, TOT_TWI, logTOT_BASIN_AREA, and logCAT_NID) as the final predictors in the stream DO model since it showed a relatively high prediction accuracy of R^2(0.59) and the lowest AIC value (77.35), compared with more complex models (Figure S15).

We also developed seasonal mean DO models (Table S2 and Figure S16). Each model had different variables in predicting the stream seasonal mean DO concentration and showed different model performance. Among the four seasonal models, winter DO had the highest accuracy (R^2=0.794) and summer DO had the lowest accuracy (R^2=0.395). The detailed equations of each model are included in the Table S2.

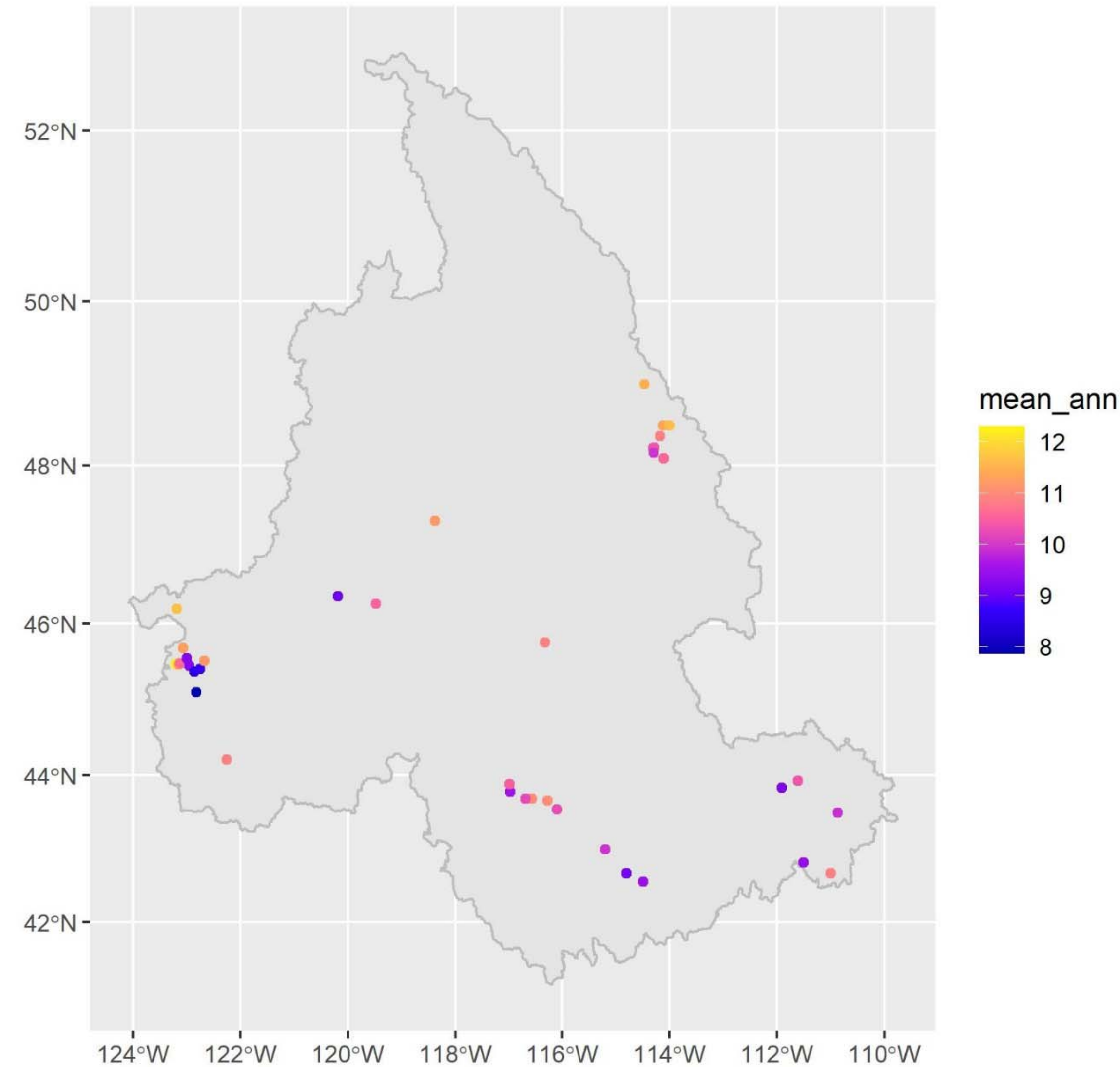

Figure S13. Temporal mean concentrations of stream DO in the CRB.

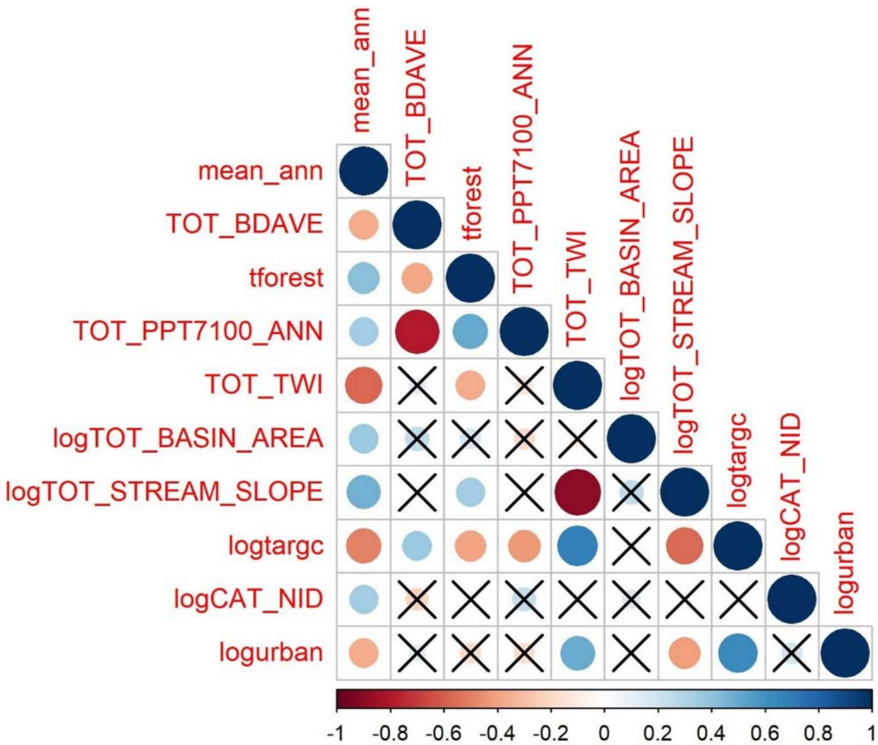

Figure S14. Spatial correlation values between mean DO concentrations and selected watershed properties.

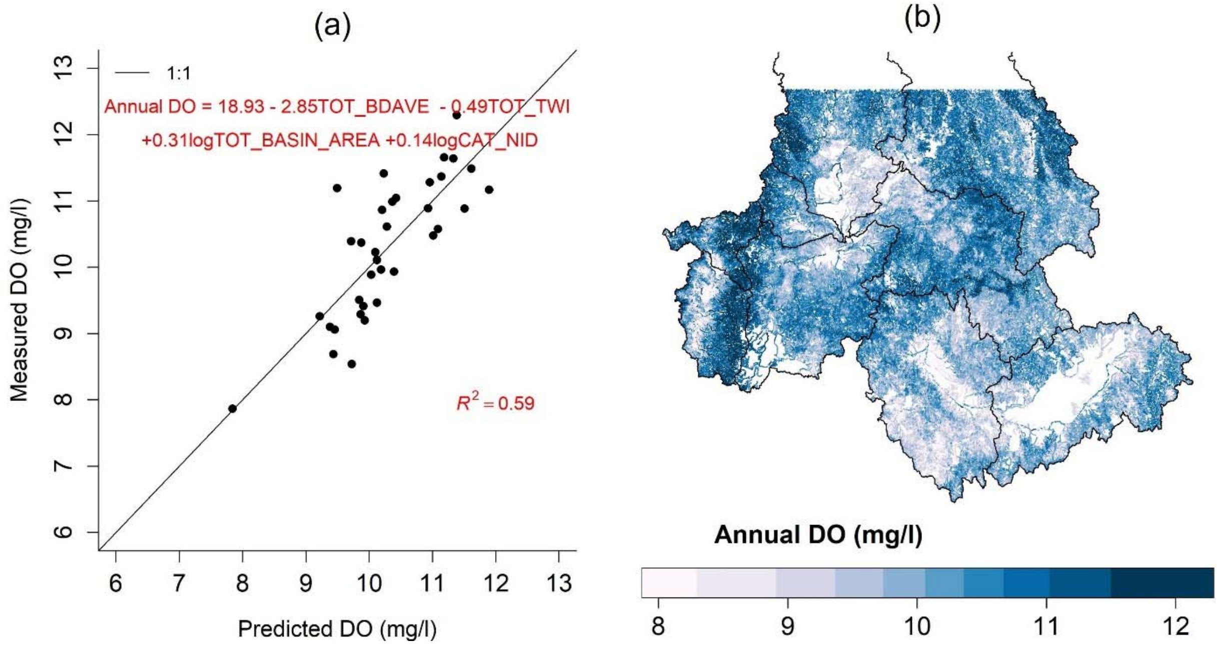

Figure S15. Developed stream DO model and its prediction: (a) developed regression model and (b) predicted stream DO concentration at the NHDPLUS stream reaches.

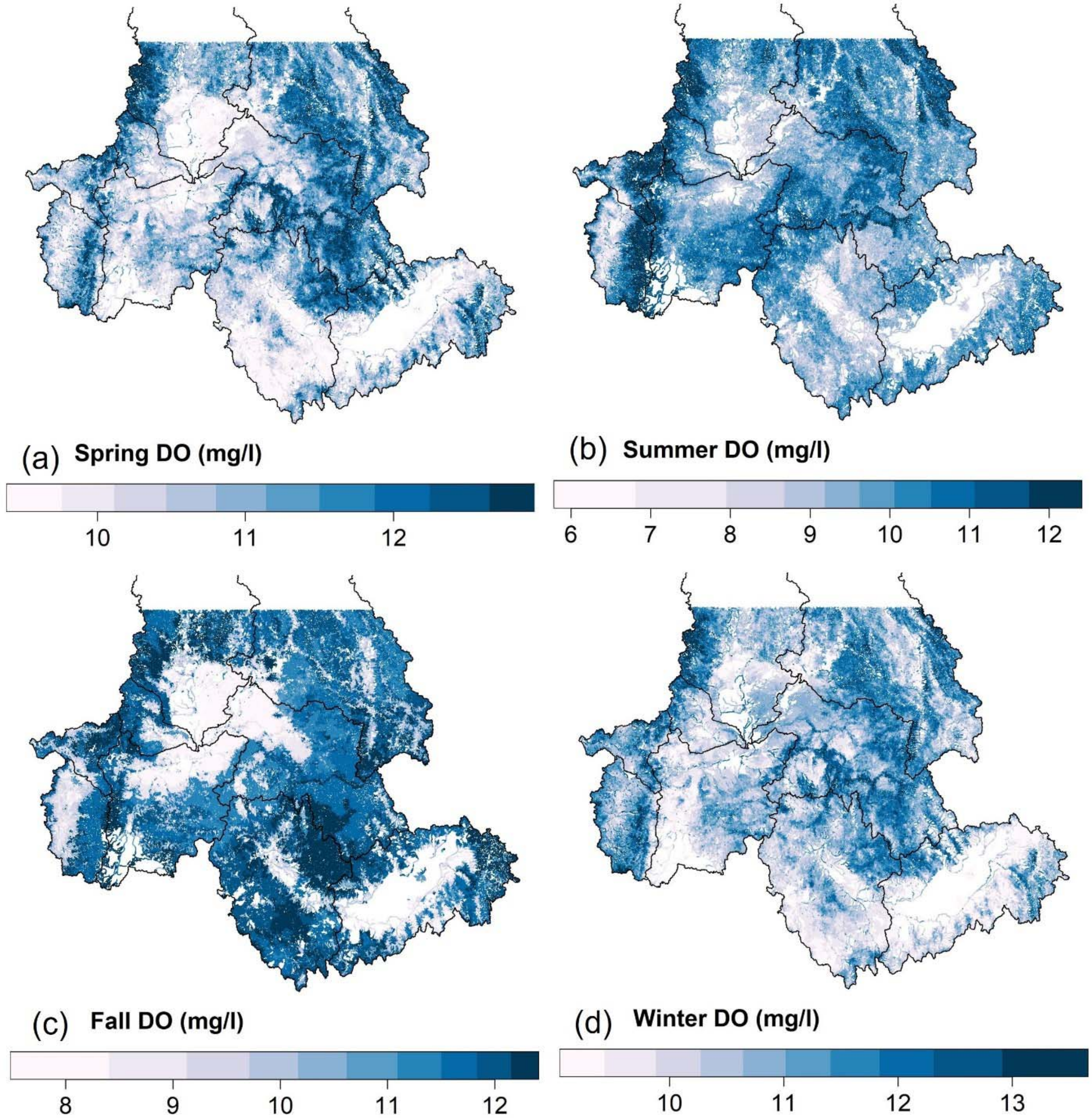

Figure S16. Seasonal stream DO models: (a) spring DO; (b) summer DO; (c) fall DO; and (d) winter DO.

For estimating the stream annual mean nitrate concentration, we used the developed 2012 SPARROW model results for the Pacific Northwest and California (Wise et al., 2019). The SPARROW model estimated the NHDPLUS-based stream flow and nutrient loading (including the stream total nitrogen, stream total phosphorus, and suspended sediment). Since our model requires a stream NO_3^- concentration, we calculated the total nitrogen concentration by dividing the total nitrogen loading with the annual streamflow estimate. Since some reaches had unrealistically high values of total nitrogen concentration due to the uncertainty of estimated flow and total nitrogen loading, we applied maximum cap values (10mg/l) to the calculated total nitrogen concentration. To test whether nitrate is a major component of total nitrogen in the stream waters, the ratio of stream nitrate concentration to the total stream nitrogen concentration was calculated for the stream gauge stations within the CRB. Figure S17 showed that the stream total nitrogen concentrations had a strong (R^2=0.99) and a linear relationship with the stream nitrate concentrations, and the median ratio of the nitrate to the total nitrogen was about 0.83. We multiplied the median ratio (0.83) to the SPARROW-based stream total nitrogen concentration to compute stream annual mean NO_3^- concentration (Figure S17c).

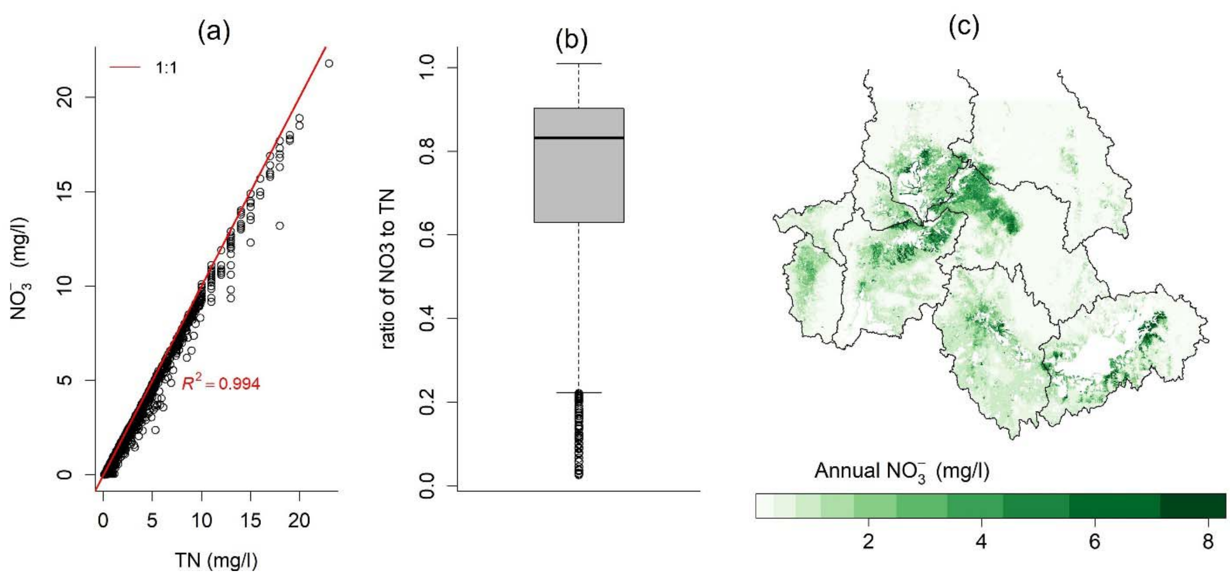

Figure S17. Prediction of stream annual mean NO_3^- concentration at the NHDPLUS stream reach scale for the CRB: (a) relationship between stream NO_3^- and stream total nitrogen concentrations at the gauge stations within the CRB; (b) ratio of the stream NO_3^- concentration to the stream total nitrogen concentration at the gauge stations within the CRB; and (c) the predicted stream NO_3^- concentration (mg/l) at the NHDPLUS stream reach scale.

Table S1. Used watershed/stream variables to build the temporal averaged stream DOC/DO model

Used variables	Variable name	Sources
Annual mean temperature (℃)	TOT_TAV7100_ANN ,CAT_TAV7100_ANN (logCAT_TAV)	PRISM,2008
Annual mean precipitation (mm)	TOT_PPT2100_ANN, CAT_PPT2100_ANN (logCAT_PPT)	PRISM,2008
Annual mean Runoff	TOT_RUN7100, CAT_RUN7100 (logCAT_RUN)	Schwarz et al., 2018
Basin drainage area (km2)	TOT_BASIN_AREA (logTOT_BASIN_AREA) CAT_BASIN_AREA	Schwarz et al., 2018
Basin elevation (m)	TOT_ELEV_MEAN (logTOT_ELEV_MEAN), CAT_ELEV_MEAN	Schwarz et al., 2018
Basin Slope	TOT_BASIN_SLOPE CAT_BASIN_SLOPE	Schwarz et al., 2018
Stream Slope	TOT_STREAM_SLOPE (logTOT_STREAM_SLOPE), CAT_STREAM_SLOPE	Schwarz et al., 2018
Soil permeability (inch/hr)	TOT_PERMAVE (logTOT_PERMAVE), CAT_PERMAVE (logCAT_PERMAVE)	STATSGO2 soil databases
Soil organic matter (%)	TOT_OM (logTOT_OM), CAT_OM	STATSGO2 soil databases
Soil bulk density(g/cm3)	TOT_BDAVE, CAT_BDAVE	STATSGO2 soil databases
% Sand	TOT_SANDAVE, CAT_SANDAVE	STATSGO2 soil databases
% Clay	TOT_CLAYAVE, CAT_CLAYAVE (logCAT_CLAYAVE)	STATSGO2 soil databases
% Silt	TOT_SILTAVE, CAT_SILTAVE	STATSGO2 soil databases
% wetland area (%)	twetland (logtwetland), wetland (logwetland)	National Land Cover Database 2001 (NLCD 2001)
% Forest area (%)	tforest, forest (logforest)	National Land Cover Database 2001 (NLCD 2001)
% Urban area (%)	turban (logturban), urban (logurban)	National Land Cover Database 2001 (NLCD 2001)
% Shrub area (%)	tshrub (logtshrub), shrub (logshrub)	National Land Cover Database 2001 (NLCD 2001)
% Agriculture area (%)	targc (logtargc) agrc (logargc)	National Land Cover Database 2001 (NLCD 2001)
Summer vegetation index	TOT_EVI_JAS_2012 (logTOT_EVI),	MODIS imagery

Used variables	Variable name	Sources
(enhanced vegetation index, EVI)	CAT_EVI_JAS_2012	
Topographic wetness index (TWI, m)	TOT_TWI, CAT_TWI	Schwarz et al., 2018
Baseflow index (BFI)	TOT_BFI, CAT_BFI	Schwarz et al., 2018
Dam storage (NID_STORAGE2010)	TOT_ NID_STORAGE2010 (logTOT_NID), CAT_NID_STORAGE2010 (logCAT_NID)	Schwarz et al., 2018
TN concentration (mg/l)	tn (logtn)	SPARROW 2012
TP concentration (mg/l)	tp (logtp)	SPARROW 2012
Parenthesis value is the variable name after log transformed. ‘CAT’ represents flowline catchment value. ‘TOT’ represents total upstream routed accumulated value. ‘tforest’ and ‘forest’ represent the percentage of combined forest lands (mixed forest, deciduous and evergreen forests) from the total upstream drainage area, and catchment drainage area, respectively. Other land classes follow the similar naming.		

Table S2. The developed seasonal stream DOC/DO models

Model	Equations	Accuracy
Spring DOC	DOC=4.56-0.03TOT_CLAYAVE-0.03tshrub-3.02CAT_EVI_JAS_2012+0.38logtargc	R^2=0.505
Summer DOC	DOC=3.11-0.02tshrub+0.44logtargc-0.16logshrub	R^2=0.359
Fall DOC	DOC=3.22-0.03tshrub+0.63logturban-0.13logshrub	R^2=0.473
Winter DOC	DOC=5.27-0.05CAT_BFI+0.47logtargc	R^2=0.54
Spring DO	DO=10.17+0.07TOT_BASIN_SLOPE+0.26logCAT_NID	R^2=0.514
Summer DO	DO=17-5.2TOT_BDAVE-0.38TOT_TWI+1.18logTOT_ELEV_MEAN	R^2=0.395
Fall DO	DO=12.4-0.05TOT_SILTAVE-0.56logtargc	R^2=0.502
Winter DO	DO=12.65+0.07TOT_BASIN_SLOPE-0.04CAT_BFI+0.08logTOT_NID+0.19logCAT_NID	R^2=0.794
‘CAT’ represents NHD flowline catchment value. ‘TOT’ represents NHD total upstream routed accumulated value. ‘tforest’ and ‘forest’ represent the percentage of combined forest lands (mixed forest, deciduous and evergreen forests) from the total upstream drainage area, and catchment drainage area, respectively. Other land classes (shrub, argc and urban) follow the similar naming. CLAYAVE: % of clay content in the soil, SILTAVE: % of silt content in the soil, BDAVE: soil bulk density, ELEV_MEAN: mean watershed’ elevation, EVI_JAS_2012: Mean enhanced vegetation Index (EVI) in summer of 2012, BASIN_SLOPE: watershed slope, TWI: topographic wetness index, BFI: Ratio of base flow to total flow and NID: Maximum dam storage between 1950 and 2010.		

Random forest model

To run the random forest model, we used the NHDPLUS version 2.1 attributes for reach catchments and modified network routed upstream watersheds for the Conterminous United States (Schwarz et al., 2018)

Table S3. Used variables in the random forest modeling for predicting hyporheic denitrification amounts in the CRB.

Variable group	Variable	Variable name	Description	Source
Climate	Annual mean temperature	CAT_TAV7100_ANN TOT_TAV7100_ANN	30-year (1971–2000) mean annual temperature (Celsius)	(McCabe & Wolock, 2016)
	Annual mean precipitation	CAT_PPT7100_ANN TOT_PPT7100_ANN	30-year (1971–2000) mean annual precipitation (mm)	(McCabe & Wolock, 2016)
Topography	Basin/catchment topography variables	TOT_BASIN_AREA TOT_BASIN_SLOPE TOT_ELEV_MEAN TOT_ELEV_MIN TOT_ELEV_MAX TOT_TWI CAT_BASIN_AREA CAT_BASIN_SLOPE CAT_ELEV_MEAN CAT_ELEV_MIN CAT_ELEV_MAX CAT_TWI	Slope, elevation maximum, and minimum and mean value, and topographic wetness index(ln(a/slope)	(Schwarz et al., 2018))
Hydrology	Annual potential evapotranspiration (PET)	TOT_PET CAT_PET	Annual averaged potential evapotranspiration(mm) from 2014–2015	(McCabe & Wolock, 2016)
	Annual actual evapotranspiration (AET)	TOT_AET CAT_AET	Annual averaged actual evapotranspiration(mm) from 2014–2015	(McCabe & Wolock, 2016)
	Annual Runoff	CAT_RUN7100 TOT_RUN7100	Estimated 30-year (1971–2000) average annual runoff	(McCabe & Wolock, 2016)
	BFI	CAT_BFI TOT_BFI	Ratio of base flow to total flow	(Schwarz et al., 2018)
	Dam storage	CAT_NID_STORAGE2010 TOT_NID_STORAGE2010	Maximum dam storage between 1950 and 2010	United States Army Corps of Engineers

Variable group	Variable	Variable name	Description	Source
Land use	% Forest area	CAT_forest TOT_forest	Deciduous/mixed and evergreen forest area	National Land Cover Database 2001 (NLCD 2001)
	% Urban area	CAT_urban TOT_urban	Developed, open Space developed, low/medium/high density area	National Land Cover Database 2001 (NLCD 2001)
	% Shrub area	CAT_shrub TOT_shrub	Dwarf scrub and Shrub/scrub	National Land Cover Database 2001 (NLCD 2001)
	% Wetland area	CAT_wetland TOT_wetland	Woody Wetlands and Emergent Herbaceous Wetlands	National Land Cover Database 2001 (NLCD 2001)
	% Agriculture	CAT_agr TOT_agr	Pasture/Hay and cultivated crops	National Land Cover Database 2001 (NLCD 2001)
	Summer vegetation index	CAT_EVI_JAS_2012 TOT_EVI_JAS_2012	Mean enhanced vegetation Index (EVI) in summer of 2012	MODIS imagery
Soil	Soil layer properties	CAT_OM TOT_OM CAT_PERMAVE TOT_PERMAVE	Soil organic matter, permeability	STATSGO2 soil databases
	Soil texture	CAT_SILTAVE CAT_CLAYAVE CAT_SANDAVE TOT_SILTAVE TOT_CLAYAVE TOT_SANDAVE	(% Silty, % CLAY and % Sand)	STATSGO2 soil databases
Stream	Contact time	CAT_CONTACT TOT_CONTACT	The length of time it takes for water to drain along subsurface flow paths to the stream	(Schwarz et al., 2018)
	Stream bankfull depth	logwbkf_m	Bankfull stream water depth	(Gomez-Velez et al., 2015)
	Stream water depth	logd_m	Stream water depth	(Gomez-Velez et al., 2015)
	Stream sinuosity	sinuosity	Flowline reach sinuosity.	(Schwarz et al., 2018)
	D50(median grain size)	D50_m	50% grain size of stream sediment materials	(Gomez-Velez et al., 2015)
	Stream slope	TOT_STREAM_SLOPE	Stream slope	

Variable group	Variable	Variable name	Description	Source
		CAT_STREAM_SLOPE		
‘CAT’ is NHD flowline catchment value, and ‘TOT’ is NHD total upstream routed accumulated value.				

References

Gomez-Velez, J. D., Harvey, J. W., Cardenas, M. B., & Kiel, B. (2015). Denitrification in the Mississippi River network controlled by flow through river bedforms. *Nature Geoscience*, *8*(12), 941–945. https://doi.org/10.1038/ngeo2567

McCabe, G. J., & Wolock, D. M. (2016). Variability and Trends in Runoff Efficiency in the Conterminous United States. *Journal of the American Water Resources Association*, *52*(5), 1046–1055. https://doi.org/10.1111/1752-1688.12431

Schwarz, G. E., Jackson, S. E., & Wieczorek, M. E. (2018). Select Attributes for NHDPlus Version 2.1 Reach Catchments and Modified Network Routed Upstream Watersheds for the Conterminous United States. *United States Geological Survey*. https://doi.org/http://dx.doi.org/10.5066/F7765D7V

Yang, Q., Zhang, X., Xu, X., & Asrar, G. R. (2017). An analysis of terrestrial and aquatic environmental controls of riverine dissolved organic carbon in the conterminous United States. *Water (Switzerland)*, *9*(6). https://doi.org/10.3390/w9060383

Zarnetske, J. P., Bouda, M., Abbott, B. W., Saiers, J., & Raymond, P. A. (2018). Generality of Hydrologic Transport Limitation of Watershed Organic Carbon Flux Across Ecoregions of the United States. *Geophysical Research Letters*, *45*(21), 11,702-11,711. https://doi.org/10.1029/2018GL080005