Supporting Information for "Prediction of Distributed River Sediment Respiration Rates using Community-Generated Data and Machine Learning"

Stefan F. Gary[2], Timothy D. Scheibe[1], Em Rexer[1],

Alvaro Vidal Torreira[2], Vanessa A. Garayburu-Caruso[1],

Amy Goldman[1], James C. Stegen[1]

[1]Pacific Northwest National Laboratory, Richland, WA, USA

[2]Parallel Works, Inc., Chicago, IL, USA

Contents of this file

1. Table S1

Additional Supporting Information (Files uploaded separately)

None

Introduction

Table S1 lists individually all the possible features (i.e. inputs) to the ML models trained in this work. This expanded table is the complement to Table 1 in the manuscript which is a summary of these features as described by overarching categories.

Table S1.

Table S1: List of features used in this manuscript. The **ID** of each feature (i.e. variable) is used to look up that feature's importance in manuscript Figure 6 across different ML models. The feature permutation importance (FPI) ratio presented in the F column corresponds to ML-run Summer-2019-log10-r08 (ID 2) in manuscript Table 2 and is presented here because it is one of only two ML models to be trained on all the features. An estimate of the uncertainty of F is presented in column s_F which is the standard deviation of F for each feature computed across the 10 SuperLearner ensemble members for the ML-run Summer-2019-log10-r08 (ID 2). The feature names from the original source data sets in the **Src** column (WH=WHONDRS, RA=RiverAtlas) are in the **Name** column, unless modified as noted in the **Description** column. Due to the presence of long feature names, spaces are inserted before underscores to automatically allow for line wrapping within sensible-width table columns. Blank spaces in the **Units** column indicate non-dimensional features.

ID	Name	F	s_F	Src	Units	Description
0	General _Vegetation	1.03	0.31	WH		General vegetation type from provided controlled vocabulary during sediment sampling. From provided controlled vocabulary with the option to select up to 2 if mixed.
1	MiniDot _Sediment	1.13	0.38	WH		Type of sediment observed at site
2	Depositional _Type	1.05	0.25	WH		Depositional zone type. From provided controlled vocabulary with option to write in other terms.
3	Hydrogeomorphology	0.95	0.13	WH		General hydrogeomorphology of river. One of the following classes may be selected: multi-channel (braided); single-channel straight; single-channel meandering.
4	Intermittent _or _Perennial	0.97	0.03	WH		Indicator of if the stream is intermittent or perennial.

Continuation of Table S1

ID	Name	F	s_F	Src	Units	Description
5	River _Gradient	1.22	0.51	WH		General gradient of river. One of the following classes may be selected: relatively flat/gentle gradient (e.g.; valleys) or relatively steep gradient (e.g.; mountainous or hilly terrain).
6	gla _pc _cse	1.00	0.00	RA	percent	Glacier extent over reach catchment.
7	Total _Heterotrophs _cells _per _gram	0.99	0.22	WH	cells/g	Total heterotrophic bacteria. Calculated by subtracting total phototrophic bacteria from total bacteria prior to rounding.
8	Total _Bacteria _cells _per _gram	0.99	0.22	WH	cells/g	Total bacteria (rounded).
9	my _lm	0.95	0.16	RA	m	Estimated length of reach based on available RA coordinates.
10	RA _lm	0.95	0.16	RA	m	Average length of reach direct from RA.
11	C _percent	1.13	0.29	WH	percent	Percent total organic carbon in <2 millimeter sediment samples (61033 US Geological Survey parameter removed from name).
12	N _percent	1.13	0.29	WH	percent	Percent total nitrogen in <2 millimeter sediment samples (01472 US Geological Survey parameter removed from name).
13	soc _th _uav	0.98	0.15	RA	tonnes/ha	Average organic carbon content in soil over all upstream watershed.
14	soc _th _cav	0.98	0.15	RA	tonnes/ha	Average organic carbon content in soil over reach catchment.
15	pac _pc _cse	0.93	0.10	RA	percent	Protected area extent over reach catchment.
16	pac _pc _use	0.93	0.10	RA	percent	Protected area extent over all upstream catchment.

Continuation of Table S1

ID	Name	F	s_F	Src	Units	Description
17	ero _kh _uav	0.96	0.14	RA	kg/ha/yr	Average soil erosion rate over all upstream watershed.
18	ero _kh _cav	0.96	0.14	RA	kg/ha/yr	Average soil erosion rate over reach catchment.
19	kar _pc _cse	1.02	0.34	RA	percent	Karst areal extent over reach catchment.
20	kar _pc _use	1.02	0.34	RA	percent	Karst areal extent over all upstream watersheds.
21	Mean _DO _mg _per _L	0.95	0.16	WH	mg/L	Dissolved oxygen measured during in situ sensor deployment.
22	Mean _DO _percent _saturation	0.98	0.13	WH	percent	Dissolved oxygen saturation measured during in situ sensor deployment.
23	crp _pc _cse	0.95	0.13	RA	percent	Cropland extent over reach catchment.
24	crp _pc _use	0.99	0.17	RA	percent	Cropland extent over all upstream watersheds.
25	slt _pc _uav	0.95	0.24	RA	percent	Silt fraction in soil over all upstream catchments.
26	slt _pc _cav	0.95	0.24	RA	percent	Silt fraction in soil over reach catchment.
27	snd _pc _uav	0.90	0.17	RA	percent	Sand fraction in soil over all upstream catchments.
28	snd _pc _cav	0.90	0.17	RA	percent	Sand fraction in soil over reach catchment.
29	cly _pc _uav	0.90	0.17	RA	percent	Clay fraction in soil over all upstream catchments.
30	cly _pc _cav	0.90	0.17	RA	percent	Clay fraction in soil over reach catchment.
31	skew _lamO2	0.91	0.14	WH		Skew (i.e. normalized 3rd central moment) of the FTICR lamO2 (feature ID 50) distribution.
32	skew _lamO20	0.98	0.17	WH		Skew (i.e. normalized 3rd central moment) of the FTICR lamO20 (feature ID 51) distribution.
33	pst _pc _use	0.96	0.14	RA	percent	Pasture extent over all upstream watersheds.
34	pst _pc _cse	0.94	0.14	RA	percent	Pasture extent over reach catchment.

Continuation of Table S1

ID	Name	F	s_F	Src	Units	Description
35	for _pc _cse	0.98	0.17	RA	percent	Forest cover extent over reach catchment.
36	for _pc _use	0.98	0.18	RA	percent	Forest cover extent over all upstream watersheds.
37	RA _ms _di	1.13	0.42	RA	m/s	Range between annual min and max average stream speed.
38	RA _SO	1.00	0.19	RA		Stream order.
39	pop _ct _usu	0.97	0.07	RA	count	Population in all upstream watersheds.
40	gdp _md _usu	0.97	0.07	RA	USD	Sum GDP over all upstream watersheds.
41	RA _cms _cyr	0.96	0.10	RA	m^3/s	Annual mean reach flow rate (originally dis _m3 _pyr).
42	RA _xam2	0.96	0.10	RA	m^2	Estimated average cross-section area derived from other RA vars.
43	RA _cms _cmx	0.96	0.10	RA	m^3/s	Annual max reach flow rate (originally dis _m3 _pmx).
44	RA _cms _cmn	0.96	0.10	RA	m^3/s	Annual min reach flow rate (originally dis _m3 _pmn).
45	RA _dm	0.96	0.10	RA	m	Estimated average depth of reach derived from other RA variables.
46	gla _pc _use	0.99	0.01	RA	percent	Glacier extent over all upstream watersheds.
47	RA _ms _av	1.07	0.30	RA	m/s	Annual average stream speed.
48	AI	0.94	0.13	WH		Aromaticity index (Koch & Dittmar, 2006, 2016).
49	DBE _O	0.98	0.18	WH		Double bond equivalent minus Oxygen (Koch & Dittmar, 2006, 2016).
50	lamO2	1.31	0.37	WH		Thermodynamic efficiency (defined by lambda) at pH 7 (Song et al., 2020).
51	lamO20	1.31	0.37	WH		Thermodynamic efficiency (defined by lambda) at standard state pH 0 (Song et al., 2020).

Continuation of Table S1

ID	Name	F	s_F	Src	Units	Description
52	delGd	1.31	0.37	WH	kJ _per _Compound _mol	Gibbs free energy of the carbon oxidation half reaction per compound mol at pH 7 (Song et al., 2020).
53	delGd0	1.31	0.37	WH	kJ _per _Compound _mol	Gibbs free energy of the carbon oxidation half reaction per compound mol at pH 0 (Song et al., 2020).
54	GFE	1.18	0.56	WH	kJ _per _C _mol	Gibbs free energy of the carbon oxidation half reaction under standard conditions (LaRowe & Van Cappellen, 2011).
55	delGcox0PerCmol	1.18	0.56	WH	kJ _per _C _mol	Gibbs free energy of the carbon oxidation half reaction under standard conditions pH 0 (LaRowe & Van Cappellen, 2011).
56	NOSC	1.18	0.56	WH		Nominal Oxidation State of Carbon (Koch & Dittmar, 2006, 2016).
57	delGcoxPerCmol	1.18	0.56	WH	kJ _per _C _mol	Gibbs free energy of the carbon oxidation half reaction at pH 7 (Song et al., 2020).
58	AI _Mod	1.01	0.29	WH		Modified aromaticity index (Koch & Dittmar, 2006, 2016).
59	perc _ConHC	1.01	0.29	WH	percent	Percent of detected chemicals this chemical class (Kim et al., 2003).
60	perc _Tannin	0.95	0.16	WH	percent	Percent of detected chemicals this chemical class (Kim et al., 2003).
61	perc _Protein	0.93	0.14	WH	percent	Percent of detected chemicals this chemical class (Kim et al., 2003).
62	perc _Lignin	0.96	0.21	WH	percent	Percent of detected chemicals this chemical class (Kim et al., 2003).
63	DBE	0.92	0.12	WH		Double bond equivalent or degree of unsaturation (Koch & Dittmar, 2006, 2016).

Continuation of Table S1

ID	Name	F	s_F	Src	Units	Description
64	perc _Lipid	0.96	0.17	WH	percent	Percent of detected chemicals this chemical class (Kim et al., 2003).
65	nli _ix _uav	1.18	0.55	RA	index	Avg. night time lights over all upstream watersheds.
66	hft _ix _u09	1.18	0.55	RA	index	Human footprint index over all upstream watersheds.
67	rdd _mk _uav	0.97	0.09	RA	m/km^2	Average road density over all upstream watersheds.
68	urb _pc _use	0.97	0.09	RA	percent	Urbanization extent over all upstream watersheds.
69	ppd _pk _uav	0.97	0.09	RA	people/km^2	Average population density over all upstream watersheds.
70	nli _ix _cav	1.24	0.47	RA	index	Average night time lights over reach catchment.
71	hft _ix _c09	1.24	0.47	RA	index	Human footprint index over reach catchment.
72	ppd _pk _cav	0.99	0.11	RA	people/km^2	Average population density over reach catchment.
73	urb _pc _cse	1.20	0.25	RA	percent	Urbanization extent over reach catchment.
74	rdd _mk _cav	0.95	0.07	RA	m/km^2	Average road density over reach catchment.
75	pop _ct _csu	0.98	0.07	RA	people	Population in reach catchment.
76	gwt _cm _cav	0.98	0.20	RA	cm	Mean ground water table depth over reach catchment.
77	slp _dg _cav	0.94	0.15	RA	degx10	Mean terrain slope over reach catchment.
78	prm _pc _use	1.03	0.06	RA	percent	Permafrost extent over all upstream watersheds.
79	prm _pc _cse	1.03	0.06	RA	percent	Permafrost extent over reach catchment.
80	tmp _dc _uyr	0.93	0.14	RA	Celsius	Annual average air temperature over all upstream watersheds.
81	tmp _dc _cyr	0.93	0.14	RA	Celsius	Annual average air temperature over reach catchment.
82	snw _pc _uyr	0.93	0.14	RA	percent	Annual average snow cover extent over all upstream watersheds.

Continuation of Table S1

ID	Name	F	s_F	Src	Units	Description
83	snw _pc _cyr	0.93	0.14	RA	percent	Annual average snow cover extent over reach.
84	snw _pc _cmx	0.93	0.15	RA	percent	Annual max snow cover extent over reach.
85	RA _lat	1.01	0.19	RA	degrees	Closest RiverAtlas site coordinates.
86	GL _lat	1.01	0.19	RA	degrees	obs/GLORICH site colocating to RA site.
87	Mean _Temp _Deg _C	0.98	0.22	WH	Celsius	Mean water temperature from dissolved oxygen (DO) sensor (miniDOT) in degrees Celsius measured during in situ sensor deployment. This calculation excludes the first 20 minutes and last 5 minutes inside the water to account for the equilibration time of the instrument for the CM Data. SSS data are trimmed to the average time of equilibrated CM data.
88	GL _lon	0.94	0.16	RA	degrees	obs/GLORICH site colocating to RA site.
89	RA _lon	0.94	0.16	RA	degrees	Closest RiverAtlas site coordinates.
90	pre _mm _cyr	0.94	0.19	RA	mm	Annual average precipitation over reach catchment.
91	pre _mm _uyr	0.94	0.19	RA	mm	Annual average precipitation over all upstream watersheds.
92	cmi _ix _cyr	0.94	0.19	RA		Annual average climate moisture index over reach catchment.
93	swc _pc _cyr	0.94	0.19	RA	percent	Annual average soil water content over reach catchment.
94	aet _mm _cdi	0.94	0.19	RA	mm	Range of annual min and max actual evaporation over reach catchment derived from RA data.

Continuation of Table S1

ID	Name	F	s_F	Src	Units	Description
95	aet _mm _cyr	0.94	0.18	RA	mm	Annual average actual evapotranspiration over reach catchment.
96	aet _mm _uyr	0.94	0.18	RA	mm	Annual average actual evaporation over all upstream watersheds.
97	swc _pc _uyr	0.91	0.14	RA	percent	Annual average soil water content over all upstream watersheds.
98	cmi _ix _uyr	0.91	0.14	RA		Annual average climate moisture index over all upstream watersheds.
99	run _mm _cyr	0.96	0.15	RA	mm/year	Annual average land surface runoff over reach catchment.
100	Percent _Silt	0.95	0.15	WH	percent	Percent silt (calculated as 100 - (percent sand + percent clay)).
101	Percent _Tot _Sand	0.91	0.21	WH	percent	Percent total sand (calculated from sum of fine; medium; and coarse sand).
102	ele _mt _cav	0.95	0.12	RA	m	Mean elevation over reach catchment.
103	tmp _dc _cdi	0.94	0.12	RA	Celsius	Range of annual min and max air temperature over reach catchment derived from RA data.
104	perc _UnsatHC	0.91	0.15	WH	percent	Percent of detected chemicals this chemical class (Kim et al., 2003).
105	Percent _Fine _Sand	1.16	0.26	WH	percent	Percent fine sand (less than 250 micrometers and greater than 53 micrometers).
106	del _15N _permil	0.91	0.12	WH	permil	Delta 15 N. Stable isotopic composition of nitrogen (delta nitrogen-15/nitrogen-14) (82338 US Geologic Survey parameter removed from name).
107	pH	0.96	0.13	WH	pH	pH.

Continuation of Table S1

ID	Name	F	s_F	Src	Units	Description
108	Water _Depth _cm	1.00	0.12	WH	centimeters	Vertical distance between water surface and bed of sampling location.
109	perc _Other	0.96	0.14	WH	percent	Percent of detected chemicals this chemical class (Kim et al., 2003).
110	cmi _ix _cdi	0.92	0.15	RA		Range of annual min and max climate moisture index derived from RA data.
111	pre _mm _cdi	0.94	0.11	RA	mm	Range of annual min and max precipitation over reach catchment derived from RA data.
112	ire _pc _cse	0.98	0.06	RA	percent	Irrigated area extent (Equipped) over reach catchment.
113	NPOC _Field _mg _per _L _as _C	1.12	0.28	WH	mg _per _L _as _carbon	Non-purgeable organic carbon from field sediment samples that were subsequently analyzed on the FTICR-MS (00681 US Geologic Survey parameter removed from name).
114	Macrophyte _Coverage	0.98	0.10	WH		Estimated macrophyte cover in the river. One of the following classes may be selected: full; partial; or no coverage.
115	perc _Carb	1.16	0.36	WH	percent	Percent of detected chemicals this chemical class (Kim et al., 2003).
116	NPOC _INC _mg _per _L _as _C	0.99	0.17	WH	mg _per _L _as _carbon	Non-purgeable organic carbon from water extractions of sediment samples that were subsequently analyzed on the FTICR-MS (00681 US Geologic Survey parameter removed from name).
117	n _chems	0.96	0.13	WH	count	Number of chemicals identified in sample from FTICR.

Continuation of Table S1

ID	Name	F	s_F	Src	Units	Description
118	Percent _Med _Sand	0.94	0.14	WH	percent	Percent medium sand (less than 500 micrometers and greater than 250 micrometers).
119	sgr _dk _rav	0.96	0.08	RA	dm/km	Avg. stream gradient over reach.
120	dor _pc _pva	0.99	0.06	RA	percent	Degree of dam regulation on reach.
121	hdi _ix _cav	0.95	0.13	RA	index	Average human development index over reach catchment.
122	perc _AminoSugar	0.91	0.12	WH	percent	Percent of detected chemicals this chemical class (Kim et al., 2003).
123	Total _Photorophs _cells _per _gram	1.08	0.12	WH	cells _per _gram	Total phototrophic bacteria (rounded).
124	Algal _Mat _Coverage	0.95	0.18	WH		Estimated algal mat cover on riverbed. One of the following classes may be selected: full; partial; or no coverage.
125	gdp _md _cav	0.96	0.15	RA	USD	Average GDP over reach catchment.
126	ire _pc _use	0.98	0.07	RA	percent	Irrigated area extent (Equipped) over all upstream watershed.
127	del _13C _permil	0.96	0.11	WH	permil	Delta 13 C. Stable isotopic composition of carbon (delta carbon-13/carbon-12) (63515 US Geological Survey parameter removed from name).
128	swc _pc _cdi	0.94	0.14	RA	percent	Range of annual min and max soil water content over reach catchment derived from RA data.

Continuation of Table S1

ID	Name	F	s_F	Src	Units	Description
129	Canopy _Cover	0.94	0.12	WH		Estimated canopy coverage over the river only. Total riparian zone canopy cover does not apply. One of the following classes may be selected: full; partial; or no coverage.
130	Percent _Clay	0.90	0.14	WH	percent	Percent clay (calculated using 1.5 hour and 24 hour hydrometer readings).
131	dist _m	0.96	0.10	RA	m	Distance between RA segment point and corresponding WHONDRS or GLORICH site.
132	Percent _Coarse _Sand	0.89	0.15	WH	percent	Percent coarse sand (less than 2000 micrometers and greater than 500 micrometers).

End of Table S1

References

Kim, S., Kramer, R. W., & Hatcher, P. G. (2003). Graphical method for analysis of ultrahigh-resolution broadband mass spectra of natural organic matter, the van krevelen diagram. *Analytical chemistry*, *75*(20), 5336–5344.

Koch, B. P., & Dittmar, T. (2006). From mass to structure: An aromaticity index for high-resolution mass data of natural organic matter. *Rapid communications in mass spectrometry*, *20*(5), 926–932.

Koch, B. P., & Dittmar, T. (2016). From mass to structure: an aromaticity index for high-resolution mass data of natural organic matter. *Rapid Communications in Mass Spectrometry*, *30*(1), 250-250. Retrieved from `https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/rcm.7433` doi: https://doi.org/

10.1002/rcm.7433

LaRowe, D. E., & Van Cappellen, P. (2011). Degradation of natural organic matter: a thermodynamic analysis. *Geochimica et Cosmochimica Acta*, *75*(8), 2030–2042.

Song, H.-S., Stegen, J. C., Graham, E. B., Lee, J.-Y., Garayburu-Caruso, V. A., Nelson, W. C., ... Scheibe, T. D. (2020). Representing organic matter thermodynamics in biogeochemical reactions via substrate-explicit modeling. *Frontiers in Microbiology*, *11*, 531756.