Machine learning models of arsenic in private wells throughout the conterminous United States as a tool for exposure assessment in human health studies

Melissa A. Lombard, Molly Scannell Bryan, Daniel K. Jones, Catherine Bulka, Paul M. Bradley, Lorraine C. Backer, Michael J. Focazio, Debra T. Silverman, Patricia Toccalino, Maria Argos, Matthew O. Gribble, Joseph D. Ayotte

Supporting Information

26 pages
16 tables
7 figures

Table SI_1. The various reporting limits for arsenic samples used in this study and the number of wells with arsenic below the reporting limits

Reporting Limit (µg/L)	Number of wells below the reporting limit
Less than 0.5	852
0.5	483
Between 0.5 and 1.0	160
1.0	4206
Between 1.0 and 2.0	164
2.0	1462
Between 2.0 and 5.0	535
5.0	686
Between 5.0 and 10	2
10	743
Total wells below a reporting limit	9293

Table SI_2. Number of wells used to develop each model

Boosted Regression Tree Models					
		Training wells		Testing wells	
Arsenic concentration threshold (µg/L)	Total wells	≤ threshold	> threshold	≤threshold	> threshold
1	16858	6125	5676	2625	2432
5	19705	11112	2682	4762	1149
10	20450	12769	1546	5472	663
Random Forest Classification Model					
Arsenic concentration category (µg/L)	Category	Training wells		Testing wells	
≤ 5	1	10947		4693	
>5 and ≤10	2	1161		498	
>10	3	1666		713	
Total Wells	All	13774		5904	

Table SI_3. Dataset descriptions and citations for all predictor variables tested in the models

Dataset	Description	units or scale	Citation or URL
Base-flow index	The percentage of streamflow attributed to baseflow at 1-km grid spacing	Percentage	Wolock, D.M., 2003, Base-flow index grid for the conterminous United States: U.S. Geological Survey Open-File Report 03–263, digital data set
Precipitation	Average annual precipitation from 1981 - 2010	inches	Thornton, P.E., M.M. Thornton, B.W. Mayer, Y. Wei, R. Devarakonda, R.S. Vose, and R.B. Cook. 2016. Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 3. ORNL DAAC, Oak Ridge, Tennessee, USA. https://doi.org/10.3334/ORNLDAAC/1328
Groundwater recharge	Average annual groundwater recharge from 1981 - 2010	inches	Hay, L.E., 2019, Application of the National Hydrologic Model Infrastructure with the Precipitation-Runoff Modeling System (NHM-PRMS), by HRU Calibrated Version: U.S. Geological Survey data release, https://doi.org/10.5066/P9NM8K8W.
Percent tile drainage	County level estimates of tile drainage	percentage	https://www.wri.org/publication/assessing-us-farm-drainage
Multi-order hydrologic position	Lateral position - the shortest distance to the stream or associated waterbody divided by the shortest distance to the stream or associated waterbody plus the shortest distance to the matching Thiessen divide.	percent x 100	Moore, R., Belitz, K., Arnold, T.L., Sharpe, J.B., and Starn, J.J., 2019, National Multi Order Hydrologic Position (MOHP) Predictor Data for Groundwater and Groundwater-Quality Modeling: U.S. Geological Survey data release, https://doi.org/10.5066/P9HLU4YY.
Soil Geochemistry	Geochemical data across the U.S. based on multi horizon sediments analyzed using a consistent set of methods.	Number, in mg/kg	Smith, D.B., Cannon, W.F., Woodruff, L.G., Solano, Federico, and Ellefsen, K.J., 2014, Geochemical and mineralogical maps for soils of the conterminous United States: U.S. Geological Survey Open-File Report 2014–1082, 386 p., https://dx.doi.org/10.3133/ofr20141082.
Surficial Geology	Unconsolidated surficial materials across the conterminous U.S.	categorical variable	Soller, D.R., Reheis, M.C., Garrity, C.P., and Van Sistine, D.R., 2009, Map database for surficial materials in the conterminous United States: U.S.

			Geological Survey Data Series 425, scale 1:5,000,000 [https://pubs.usgs.gov/ds/425/].
General Lithology	Generalized rock types across the conterminous U.S.	categorical variable	Anning, D.W. and Ator, S.W., 2017, Generalized lithology for the conterminous United States: U.S. Geological Survey data release, https://doi.org/10.5066/F7R78D4N.
Bedrock Geology	Bedrock geology based on King and Beikman 1974 map	categorical variable	Schruben, P.G., Arndt, R.E., Bawiec, W.J., 1997, Geology of the Conterminous United States at 1:2,500,000 Scale--A Digital Representation of the 1974 P.B. King and H.M. Beikman Map: U.S. Geological Survey Digital Data Series DDS-11 release 2, 26 p. https://pubs.usgs.gov/dds/dds11/
Ecoregion	Ecoregions denote areas of general similarity in ecosystems and in the type, quality, and quantity of environmental resources.	categorical variable	US Environmental Protection Agency, 2013, Level III Ecoregions of the Conterminous United States, U.S. EPA Office of Research and Development (ORD) - National Health and Environmental Effects Research Laboratory (NHEERL).

Table SI_4. Predictor variable definitions included in the final models.

Variable	Variable Definition	Source Dataset
as_c	arsenic concentration in the soil C horizon	Soil Geochemistry
be_c	beryllium concentration in the soil C horizon	Soil Geochemistry
BFI	percent of total stream flow that is estimated to be base-flow	Base-flow index
bi_c	bismuth concentration in the soil C horizon	Soil Geochemistry
D3	Devonian; stratified sequence, mainly marine, most complete sequences; Upper Devonian	Bedrock Geology
De	Devonian eugeosynclinal deposits	Bedrock Geology
DMppt8110	Average annual precipitation from 1981 to 2010	Precipitation
DSe	New England; Devonian and Silurian eugeosynclinal deposits	Bedrock Geology
inorgc_c	inorganic carbon concentration in the soil C horizon	Soil Geochemistry
Kg	Plutonic and intrusive rocks; Cretaceous granitic rocks; Western United States	Bedrock Geology
LP2	Lateral position for second order streams	Multi-order hydrologic position
LP4	Lateral position for fourth order streams	Multi-order hydrologic position
LP6	Lateral position for sixth order streams	Multi-order hydrologic position
mo_c	molybdenum concentration in the soil C horizon	Soil Geochemistry
na_6.2	Western Cordillera ecoregion	Ecoregion
na_7.1	Marine West Coast forest ecoregion	Ecoregion
na_8.1	Mixed wood plains ecoregion	Ecoregion

na_8.2	Central USA plains ecoregion	Ecoregion
na_9.2	Temperate prairies ecoregion	Ecoregion
na_9.3	West-Central semi-arid prairies ecoregion	Ecoregion
na_9.4	South Central semi-arid prairies ecoregion	Ecoregion
na_9.5	Texas-Louisiana coastal plain ecoregion	Ecoregion
na_9.6	Tamaulipas-Texas semiarid plain ecoregion	Ecoregion
na_10.1	Cold deserts ecoregion	Ecoregion
na_11.1	Mediterranean California ecoregion	Ecoregion
na_12.1	Western Sierra Madre piedmont ecoregion	Ecoregion
na_13.1	Upper Gila Mountains ecoregion	Ecoregion
ni_c	Nickel concentration in the soil C horizon	Soil Geochemistry
O2	Ordovician; stratified sequence, mainly marine, most complete sequences; Middle Ordovician (Mohawkian), Upper Ordovician included in places; Eastern United States	Bedrock Geology
Oe	Ordovician eugeosynclinal deposits; Appalachian region	Bedrock Geology
orgc_c	organic carbon concentration in the soil C horizon	Soil Geochemistry
p_c	phosphorus concentration in the soil C horizon	Soil Geochemistry
Percent_Ti	Percent of area by county with tile drainage	Percent tile drainage
PP4	Upper Pennsylvanian; stratified sequence mainly marine, most complete sequences; Virgilian Series (Monongahela Group); Central interior region, from Appalachians to Kansas and central Texas	Bedrock Geology
PRMS8110Re	Average annual groundwater recharge from 1981 to 2010	Groundwater Recharge
Pzg1	Plutonic and intrusive rocks; Lower Paleozoic granitic rocks; about 400 - 500 my. In New England, Highlandcroft and Oliverian Plutonic Series, related to Taconian orogeny; in Piedmont province premetamorphic; Appalachian region (New England and Piedmont province), rarely in western United States.	Bedrock Geology
Pzg2	Plutonic and intrusive rocks; Middle Paleozoic granitic rocks; about 350-400 my. In New England, New Hampshire Plutonic Series and others, related to Acadian orogeny; in Piedmont province, synmetamorphic; Appalachian region (New England and Piedmont province), rarely in western United States	Bedrock Geology
Q	Quaternary; stratified sequence, mainly marine, combinations; Western United States; only large areas and thick deposits shown	Bedrock Geology
Qv	Quaternary volcanic rocks	Bedrock Geology
rt_carb	Generalized lithology classified as CARB: sediments and sedimentary rocks. Carbonate rocks such as limestone and dolostone. Generally, any rock including any minor carbonate lithology is included in this group. Some special cases exist where carbonates are also identified based on LITH62MINO field in Schweitzer, Peter N., 2011, Combined geologic map data for the conterminous US derived from the USGS state geologic map compilation.	General Lithology

rt_clast_c	Generalized lithology classified as CLAST_C: sediments and sedimentary rocks. Clastic sediments/rocks primarily made of sands, gravels, cobbles, or larger clasts.	General Lithology
rt_clast_f	Generalized lithology classified as CLAST_F: sediments and sedimentary rocks. Clastic sediments/rocks primarily made of fine-grained materials such as shale, siltstone, claystone, mudstone, etc.	General Lithology
rt_clast_u	Generalized lithology classified as CLAST_U: sediments and sedimentary rocks. Clastic sediments/rocks of unknown or highly variable clast sizes.	General Lithology
rt_meta	Generalized lithology classified as META: Metamorphic rocks.	General Lithology
rt_plut_qtz	Generalized lithology classified as PLUT_QTZ: Igneous, generally felsic, quartz-rich plutonic rocks such as granitoids, granite, granodacite. Lithology of plutonics and volcanics are split out based on quartz content using the IUGS classification presented in Hyndman's 'Petrology of Igneous and Metamorphic Rocks' second edition, 1985, page 33, fields 1-5.	General Lithology
S2	Silurian; stratified sequence, mainly marine, most complete sequences; Middle Silurian (Niagaran)	Bedrock Geology
sb_c	antimony concentration in the soil C horizon	Soil Geochemistry
Se	Silurian eugeosynclinal deposits	Bedrock Geology
se_c	selenium concentration in the soil C horizon	Soil Geochemistry
Te2	Tertiary; Eocene; stratified sequence, mainly marine, most complete sequences; Claiborne Group Atlantic and Gulf Coastal Plains	Bedrock Geology
Tm	Tertiary; stratified sequence, mainly marine, most complete sequences; Miocene, includes Pliocene in Southeastern U.S.	Bedrock Geology
Toc	Tertiary; Oligocene continental deposits	Bedrock Geology
Tmc	Tertiary; Miocene continental deposits	Bedrock Geology
Tp	Tertiary; stratified sequence, mainly marine, most complete sequences; Pliocene	Bedrock Geology
Tpc	Tertiary; Pliocene continental deposits includes Miocene and Quaternary deposits in places	Bedrock Geology
Tpv	Tertiary; Pliocene volcanic rocks	Bedrock Geology
Tr	Triassic; stratified sequence, mainly marine, most complete sequences; mainly marine but includes continental Dockum Group in Texas and New Mexico, and Newark Group in Appalachian region	Bedrock Geology
Txc	Tertiary; Paleocene continental deposits	Bedrock Geology
uK2	stratified sequence, mainly marine, most complete sequences; Upper Cretaceous Austin and Eagle Ford Groups (=Colorado Group); Atlantic and Gulf Coastal Plains and western interior region	Bedrock Geology
uK3	stratified sequence, mainly marine, most complete sequences; Upper Cretaceous Taylor Group (= Montana Group), In Montana plains, separated into uK3a and b at	Bedrock Geology

	base of Judith River Formation; Atlantic and Gulf Coastal Plains and western interior region	
uK3b	stratified sequence, mainly marine, most complete sequences; Upper Cretaceous Taylor Group (= Montana Group), In Montana plains, separated into uK3a and b at base of Judith River Formation; Atlantic and Gulf Coastal Plains and western interior region	Bedrock Geology
uK4	stratified sequence, mainly marine, most complete sequences; Upper Cretaceous Navarro Group (Laramie and other formations); Atlantic and Gulf Coastal Plains and western interior region	Bedrock Geology
uc_12	Alluvial sediments, thick unit thickness > 100 feet, geologic age Holocene to Pliocene	Surficial Geology
uc_221	Coastal zone sediments, mostly medium-grained, unit thickness <100 feet, geologic age Holocene to Pliocene	Surficial Geology
uc_312	Eolian sediments, mostly loess, thick, thickness >100 feet, geologic age Holocene to Pliocene	Surficial Geology
uc_321	Eolian sediments, mostly dune sand, thin, unit thickness <100 feet, geologic age Holocene to Pleistocene	Surficial Geology
uc_322	Eolian sediments, mostly dune sand, thick, thickness >100 feet, geologic age Holocene to Pleistocene	Surficial Geology
uc_331	Eolian sediments on southern High Plains, unit thickness <100 feet, geologic age Holocene to Pleistocene	Surficial Geology
uc_421	Glacial till sediments, mostly silty, thin, unit thickness <100 feet, geologic age late Wisconsinan to pre-Illinoian	Surficial Geology
uc_422	Glacial till sediments, mostly silty, thick, unit thickness >100 feet, geologic age late Wisconsinan to pre-Illinoian	Surficial Geology
uc_431	Glacial till sediments, mostly sandy, think, unit thickness <100 feet, geologic age late Wisconsinan to pre-Illinoian	Surficial Geology
uc_451	Glaciofluvial ice-contact sediments, mostly sand and gravel, thin, unit thickness <100 feet, geologic age late Wisconsinan to Illinoian	Surficial Geology
uc_620	Colluvial and alluvial sediments, discontinuous, or patchy in distribution, geologic age Holocene to Tertiary	Surficial Geology
uc_811	Proglacial sediments, mostly fine grained, thin, unit thickness <100 feet, geologic age late Wisconinan to pre-Illinoian	Surficial Geology
uc_812	Proglacial sediments, mostly fine grained, thin, unit thickness >100 feet, geologic age late Wisconinan to pre-Illinoian	Surficial Geology
uc_821	Proglacial sediments, mostly coarse-grained, thin, unit thickness <100 feet, geologic age late Wisconsinan to pre-Illinoian	Surficial Geology
uc_822	Proglacial sediments, mostly coarse-grained, thick, unit thickness >100 feet, geologic age late Wisconsinan to pre-Illinoian	Surficial Geology
uc_910	Residual materials developed in igneous and metamorphic rocks, unit thickness discontinuous, or patchy in distribution, geologic age Holocene to Tertiary	Surficial Geology

uc_920	Residual materials developed in sedimentary rocks, discontinuous, unit thickness discontinuous, or patchy in distribution, geologic age Holocene to Tertiary	Surficial Geology
uc_940	Residual materials developed in carbonate rocks, discontinuous, unit thickness discontinuous, or patchy in distribution, geologic age Holocene to Tertiary	Surficial Geology
uc_960	Residual materials developed in bedrock, with alluvial sediments, discontinuous, unit thickness discontinuous, or patchy in distribution, geologic age Holocene to Tertiary	Surficial Geology
uc_970	Residual materials developed in bedrock, discontinuous, unit thickness discontinuous, or patchy in distribution, geologic age Holocene to Tertiary	Surficial Geology
uc_999	water	Surficial Geology
uPz	stratified sequence, mainly marine, combinations; Upper Paleozoic; Western United States	Bedrock Geology
Wgn	PreCambrian; metamorphic rocks; Orthogneiss and paragneiss	Bedrock Geology

Table SI_5. Predictor variables and their relative influence in the final models

BRT1		BRT5		BRT10		RFC	
Variable	Relative Influence	Variable	Relative Influence	Variable	Relative Influence	Variable	Relative Influence
DMppt8110	15.32	DMppt8110	13.51	DMppt8110	16.48	DMppt8110	11.29
se_c	7.16	as_c	5.79	p_c	6.00	LP6	8.45
orgc_c	4.79	LP6	5.65	as_c	5.86	LP4	7.97
p_c	4.72	PRMS8110Re	5.30	BFI	5.03	LP2	7.96
PRMS8110Re	4.64	p_c	5.24	se_c	4.91	PRMS8110Re	5.14
sb_c	4.25	se_c	5.01	sb_c	4.88	as_c	4.96
mo_c	3.99	orgc_c	4.85	orgc_c	4.20	BFI	4.96
be_c	3.50	BFI	4.64	be_c	4.06	p_c	4.74
LP6	3.49	LP4	4.63	LP6	4.05	sb_c	4.64
as_c	3.49	LP2	4.43	PRMS8110Re	4.04	se_c	4.38
bi_c	3.33	be_c	4.41	LP4	3.93	mo_c	4.03
inorgc_c	3.32	sb_c	4.06	mo_c	3.81	orgc_c	4.02
LP4	3.28	inorgc_c	3.93	LP2	3.50	be_c	4.01
BFI	3.28	mo_c	3.92	inorgc_c	3.49	inorgc_c	3.96
Percent_Ti	3.19	ni_c	3.83	ni_c	3.47	ni_c	3.70
ni_c	3.19	bi_c	3.55	Percent_Ti	3.34	bi_c	3.54
LP2	2.89	Percent_Ti	2.67	bi_c	2.91	Tpc	1.58
na_8.1	2.59	Tpc	1.79	na_8.1	1.87	Percent_Ti	1.50
Tpc	2.41	na_8.1	1.51	Q	1.28	Q	1.15
Q	1.91	Tpv	0.97	Tpc	1.20	na_8.1	0.89
uK3	1.50	Q	0.97	uc_811	1.20	uc_331	0.71

na_6.2	0.86	uc_331	0.95	uc_331	1.01	Tpv	0.58
Tr	0.67	Se	0.78	uc_431	0.96	rt_clast_c	0.52
uc_920	0.64	rt_clast_c	0.50	Tpv	0.90	rt_clast_f	0.47
uK2	0.63	uc_431	0.47	Se	0.79	uc_12	0.39
uK4	0.58	Tp	0.39	DSe	0.69	rt_clast_u	0.39
rt_carb	0.54	rt_clast_f	0.38	uc_12	0.63	Se	0.39
na_9.5	0.53	uc_312	0.36	Tp	0.60	Tp	0.35
uc_12	0.52	uc_811	0.36	uc_812	0.58	uc_811	0.31
uc_910	0.49	DSe	0.35	uc_422	0.54	rt_meta	0.27
na_9.3	0.49	uc_812	0.35	uc_322	0.52	na_10.1	0.27
rt_clast_f	0.43	uc_12	0.34	na_10.1	0.46	rt_plut_qtz	0.26
uc_331	0.42	rt_carb	0.34	Pzg2	0.39	rt_carb	0.25
Txc	0.38	uK3b	0.33	rt_carb	0.36	na_13.1	0.22
Tpv	0.34	uc_422	0.31	uc_821	0.35	DSe	0.22
uc_422	0.34	uc_999	0.28	uc_970	0.34	uc_431	0.22
DSe	0.33	Pzg2	0.27	S2	0.31	uc_321	0.20
Se	0.30	uc_322	0.26	uc_451	0.28	na_9.4	0.15
uc_451	0.29	uc_821	0.25	De	0.28	na_6.2	0.13
uc_822	0.27	uc_221	0.23	uc_321	0.24	na_11.1	0.13
Qv	0.25	S2	0.23	na_13.1	0.24	na_9.2	0.11
na_7.1	0.25	uc_970	0.22			na_8.2	0.10
Tm	0.24	rt_clast_u	0.22			Tmc	0.10
uc_940	0.24	uc_321	0.22			Wgn	0.09
D3	0.24	Toc	0.21			Oe	0.09
rt_plut_qtz	0.23	na_9.5	0.21			PP4	0.08
na_12.1	0.22	Pzg1	0.21			na_7.1	0.08
uc_970	0.21	na_10.1	0.19			uc_620	0.05
uc_421	0.21	na_7.1	0.17				
O2	0.20						
uc_821	0.20						
na_11.1	0.19						
Tp	0.18						
Pzg1	0.18						
na_9.6	0.17						
rt_clast_u	0.16						
Kg	0.16						
De	0.16						
uc_960	0.15						
uc_811	0.15						
uPz	0.15						
Te2	0.15						
Toc	0.15						
uc_431	0.14						
rt_meta	0.13						

Table SI_6. Regional classifications based on the USGS Ground Water Atlas of the United States

USGS ground water atlas region	States within the region
B	California, Nevada
C	Arizona, Colorado, New Mexico, Utah
D	Kansas, Missouri, Nebraska
E	Oklahoma, Texas
F	Arkansas, Louisiana, Mississippi
G	Alabama, Florida, Georgia, South Carolina
H	Idaho, Oregon, Washington
I	Montana, North Dakota, South Dakota, Wyoming
J	Iowa, Michigan, Minnesota, Wisconsin
K	Illinois, Indiana, Kentucky, Ohio, Tennessee
L	Delaware, Maryland, New Jersey, North Carolina, Pennsylvania, Virginia, West Virginia
M	Connecticut, Maine, Massachusetts, New Hampshire, New York, Rhode Island, Vermont

Table SI_7. Regional results for BRT 1 training data

Region	Total wells	Wells < 1 µg/L	Wells >1 µg/L	Accuracy	Specificity	Sensitivity	Kappa
B	1152	382	770	0.8333	0.5838	0.9571	0.5899
C	1182	627	555	0.8621	0.8868	0.8342	0.7226
D	418	197	221	0.8373	0.8325	0.8416	0.6737
E	1589	710	879	0.8974	0.9479	0.8567	0.795
F	248	210	38	0.9032	0.9952	0.3947	0.5112
G	220	182	38	0.8773	0.9890	0.3421	0.4354
H	1850	726	1124	0.8643	0.8003	0.9057	0.7127
I	976	628	348	0.8719	0.9475	0.7356	0.7099
J	784	358	426	0.861	0.8184	0.8967	0.7185
K	326	169	157	0.8804	0.8817	0.8790	0.7605
L	1559	1190	369	0.8711	0.9714	0.5474	0.5925
M	1497	746	751	0.835	0.8391	0.8309	0.67

Table SI_8. Regional results for BRT1 testing data

Region	Total wells	Wells < 1 µg/L	Wells >1 µg/L	Accuracy	Specificity	Sensitivity	Kappa
B	498	170	328	0.7309	0.4235	0.8902	0.344
C	502	270	232	0.7629	0.7815	0.7414	0.523
D	199	85	114	0.7739	0.6941	0.8333	0.533
E	702	310	392	0.8561	0.9032	0.8189	0.7123
F	123	109	14	0.878	0.98165	0.07143	0.0807
G	74	64	10	0.8919	0.9688	0.4000	0.4436
H	782	327	455	0.7404	0.6514	0.8044	0.4607
I	392	242	150	0.773	0.8967	0.5733	0.4945
J	326	153	173	0.7209	0.6340	0.7977	0.4351
K	147	73	74	0.7619	0.8219	0.7027	0.5242
L	667	514	153	0.8201	0.9514	0.3791	0.3937
M	645	308	337	0.6992	0.7565	0.6469	0.4009

Table SI_9. Regional results for BRT5 training data

Region	Total wells	Wells < μg/L	Wells >5 μg/L	Accuracy	Specificity	Sensitivity	Kappa
B	1171	826	345	0.9163	0.9855	0.7507	0.7852
C	1221	907	314	0.9304	0.9857	0.7707	0.8059
D	642	514	128	0.9003	0.9883	0.5469	0.6315
E	2455	2076	379	0.9511	0.9908	0.7335	0.7946
F	273	264	9	0.967	1.00	0	0
G	233	221	12	0.97	0.9955	0.500	0.617
H	1883	1407	476	0.9076	0.9787	0.6975	0.7343
I	1010	856	154	0.9287	0.9907	0.5844	0.6758
J	1079	867	212	0.8906	0.9954	0.4623	0.5692
K	415	324	91	0.9036	0.9722	0.6593	0.6917
L	1809	1671	138	0.9375	1.0000	0.1812	0.2901
M	1603	1179	424	0.8927	0.9763	0.6604	0.6978

Table SI_10. Regional results for BRT5 testing data

Region	Total wells	Wells < 5 μg/L	Wells >5 μg/L	Accuracy	Specificity	Sensitivity	Kappa
B	489	347	142	0.8098	0.9164	0.5493	0.5023
C	503	387	116	0.8827	0.9793	0.5603	0.6202
D	287	233	54	0.8328	0.9700	0.2407	0.2779
E	1127	927	200	0.9139	0.9752	0.6300	0.6724
F	103	100	3	0.9709	1.000	0.000	0
G	89	85	4	0.9326	0.9765	0.0000	-0.0309
H	752	558	194	0.8112	0.9355	0.4536	0.441
I	442	381	61	0.9005	0.9816	0.3934	0.4727
J	442	359	83	0.8213	0.9833	0.1205	0.1505
K	211	170	41	0.872	0.9588	0.5122	0.5354
L	753	701	52	0.9309	1.0000	0.0000	0
M	713	514	199	0.7658	0.9241	0.3568	0.3255

Table SI_11. Regional results for BRT10 training data

Region	Total wells	Wells < 10 μg/L	Wells > 10 μg/L	Accuracy	Specificity	Sensitivity	Kappa
B	1165	973	192	0.9253	0.9949	0.5729	0.6767
C	1207	999	208	0.9445	0.9910	0.7212	0.7854
D	674	639	35	0.9481	1.00	0	0
E	2951	2749	202	0.9648	0.9964	0.5347	0.6577
F	267	264	3	0.9888	1.00	0	0
G	230	227	3	0.987	1.00	0	0
H	1796	1521	275	0.9109	0.9862	0.4945	0.5832
I	1043	951	92	0.9521	0.9979	0.4783	0.615
J	1101	992	109	0.931	0.9970	0.3303	0.4582
K	437	377	60	0.9245	0.9920	0.5000	0.6069
L	1839	1777	62	0.9663	1.0	0	0
M	1605	1300	305	0.8966	0.9854	0.5180	0.5997

Table SI_12. Regional results for BRT10 testing data

Region	Total wells	Wells < 10 µg/L	Wells > 10 µg/L	Accuracy	Specificity	Sensitivity	Kappa
B	495	408	87	0.8929	0.9706	0.5287	0.5747
C	521	442	79	0.9136	0.9774	0.5570	0.6141
D	263	247	16	0.9392	1	0	0
E	1274	1179	95	0.9411	0.9830	0.4211	0.4865
F	109	109	0	1	1	NA	NA
G	93	92	1	0.9892	1	0	0
H	840	720	120	0.8881	0.9750	0.3667	0.4278
I	411	371	40	0.9294	0.9865	0.4000	0.4904
J	459	418	41	0.915	0.99522	0.09756	0.1508
K	189	172	17	0.8995	0.9709	0.1765	0.1936
L	766	725	41	0.9465	1	0	0
M	715	589	126	0.8224	0.9508	0.2222	0.2204

Table SI_13. Regional results for RFC testing data

Region	Total wells	C1 wells	C2 wells	C3 wells
B	514	344	68	102
C	526	385	43	98
D	262	205	40	17
E	1050	874	79	97
F	110	103	4	3
G	116	109	2	5
H	786	571	95	120
I	432	360	23	49
J	470	386	42	42
K	197	162	17	18
L	742	679	32	31
M	699	515	53	131

Table SI_14. Regional results for RFC testing data

Region	Overall Accuracy	Kappa	User's Accuracy C1	User's Accuracy C2	User's Accuracy C3	Producer's Accuracy C1	Producer's Accuracy C2	Producer's Accuracy C3
B	0.7510	0.4031	0.948	0.103	0.520	0.773	0.412	0.707
C	0.8118	0.4969	0.948	0.140	0.571	0.853	0.273	0.737
D	0.7748	0.1775	0.951	0.150	0.118	0.812	0.500	0.200
E	0.8752	0.5227	0.968	0.241	0.557	0.922	0.413	0.628
F	0.9363	0	1	0	0	0.936	NA	NA
G	0.9397	0.2039	0.991	0	0.200	0.947	NA	0.50
H	0.7595	0.3683	0.932	0.168	0.408	0.829	0.32	0.521
I	0.875	0.434	0.983	0.043	0.469	0.885	0.167	0.885
J	0.8085	0.2255	0.938	0.071	0.357	0.850	0.273	0.455
K	0.8426	0.4103	0.944	0.118	0.611	0.884	0.400	0.579
L	0.9084	0.0745	0.988	0	0.097	0.919	0	0.30
M	0.7654	0.3246	0.93	0	0.427	0.805	0	0.629

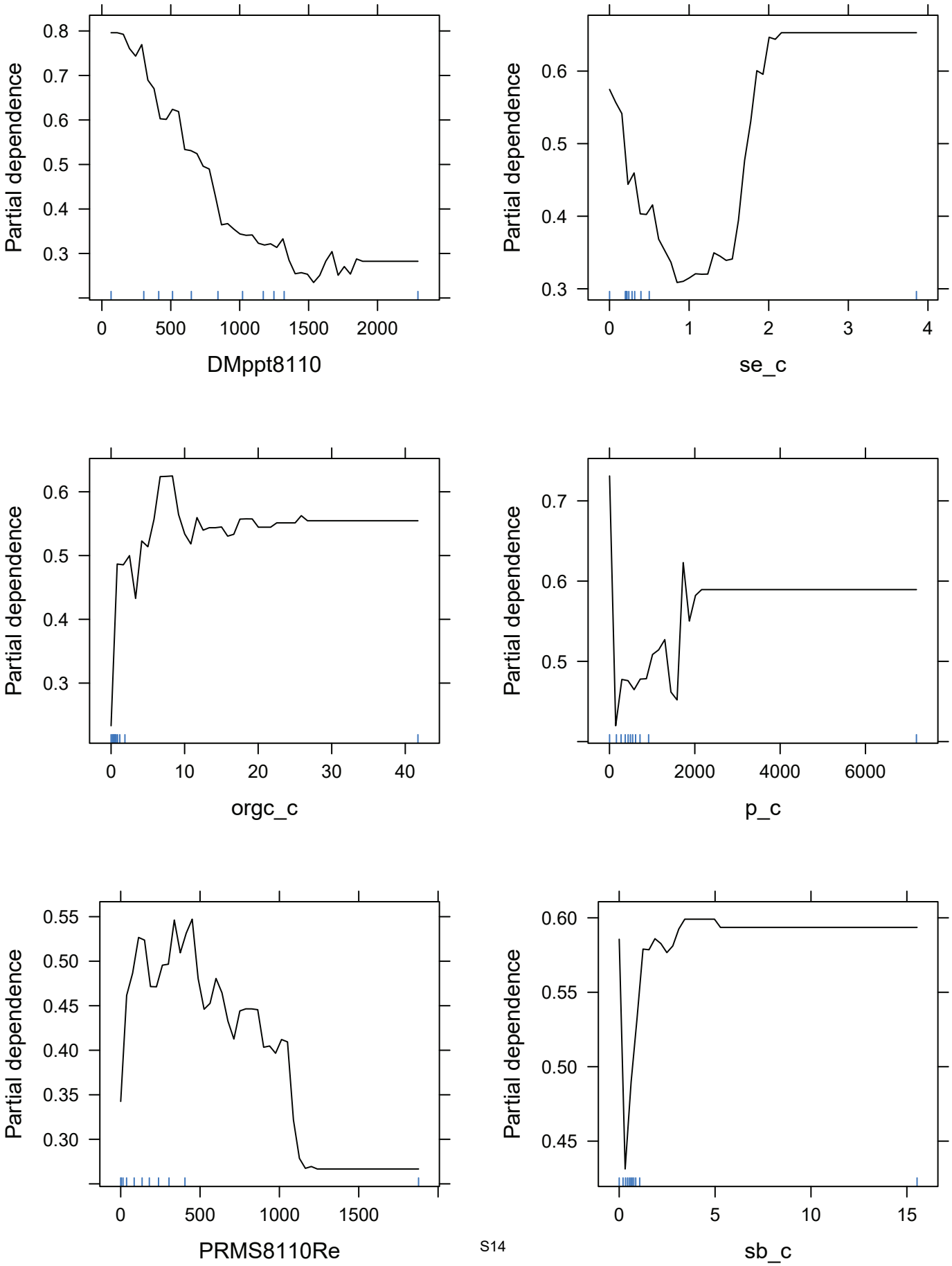
Table SI_15. Regional results for RFC training data

Region	Total wells	C1 wells	C2 wells	C3 wells
B	1145	780	155	210
C	1195	875	108	212
D	667	528	84	55
E	2532	2117	201	214
F	266	260	5	1
G	206	195	5	6
H	1831	1337	196	298
I	1020	865	65	90
J	1049	831	111	107
K	429	319	39	71
L	1820	1672	61	87
M	1614	1168	131	315

Table SI_16. Regional results for RFC training data

Region	Overall Accuracy	Kappa	User's Accuracy C1	User's Accuracy C2	User's Accuracy C3	Producer's Accuracy C1	Producer's Accuracy C2	Producer's Accuracy C3
B	1	1	1	1	1	1	1	1
C	1	1	1	1	1	1	1	1
D	1	1	1	1	1	1	1	1
E	1	1	1	1	1	1	1	1
F	1	1	1	1	1	1	1	1
G	1	1	1	1	1	1	1	1
H	1	1	1	1	1	1	1	1
I	1	1	1	1	1	1	1	1
J	1	1	1	1	1	1	1	1
K	0.9977	0.9943	1	1	0.986	0.997	1	1
L	0.9995	0.9964	1	0.984	1	0.999	1	1
M	1	1	1	1	1	1	1	1

Figure SI_1. Partial Dependence Plots for the 10 most influential variables in the BRT1 model. Variables are defined in Table SI_2. Blue hatch lines on the x-axis indicate the deciles of the variable in the model training dataset.



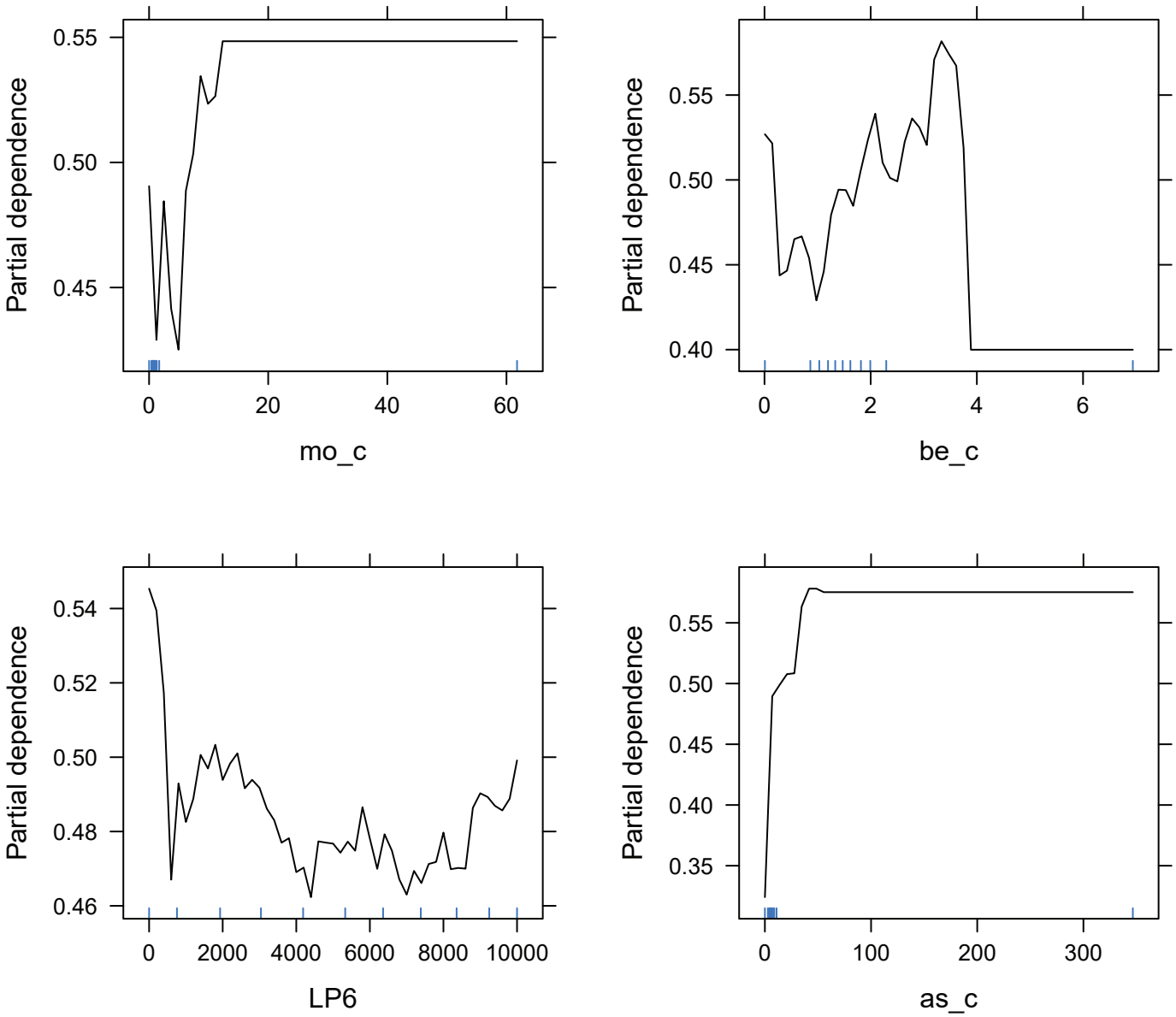
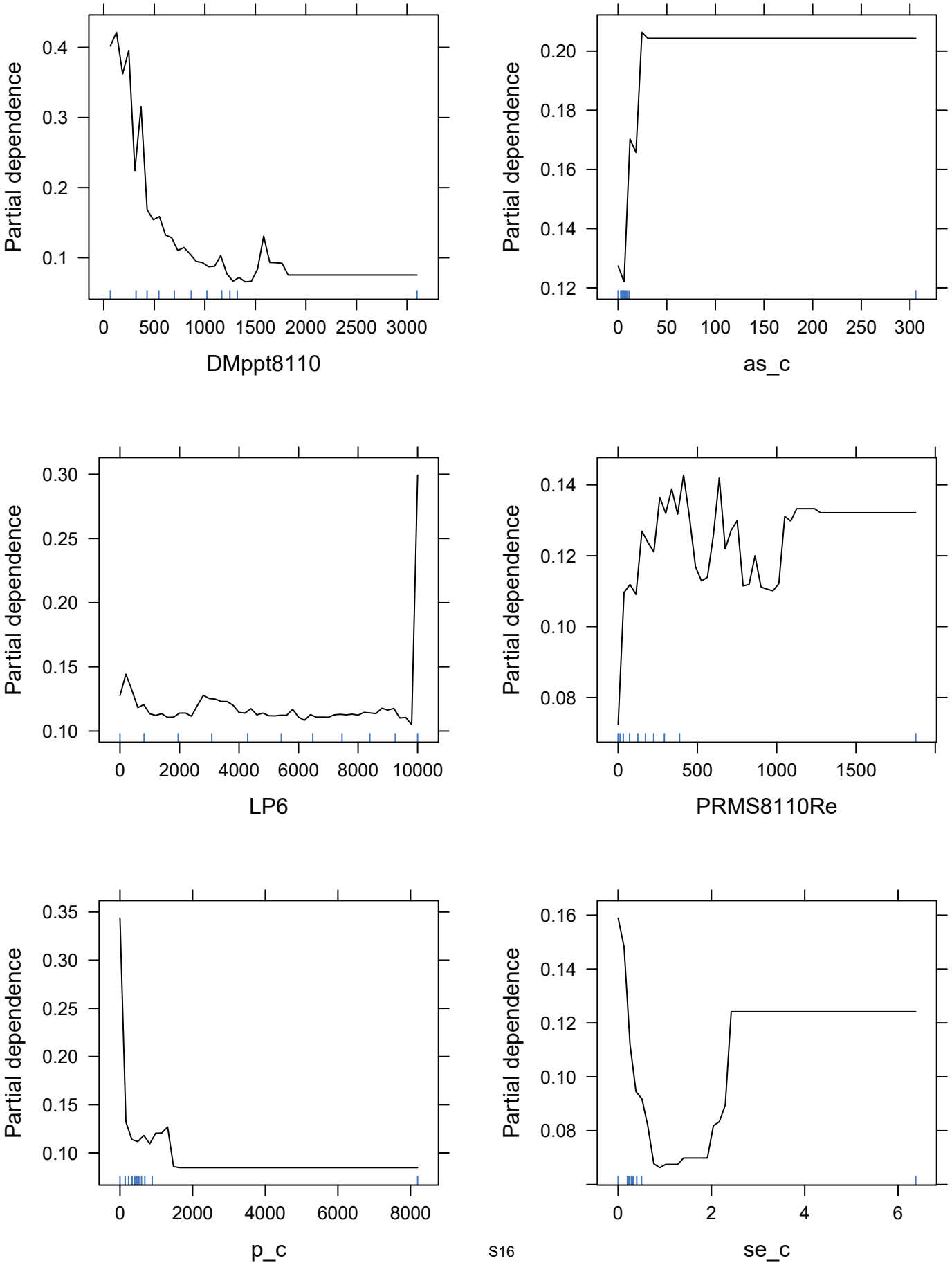

Partial dependence
0.55
0.50
0.45
0
20
40
60
mo_c
Partial dependence
0.55
0.50
0.45
0.40
0
2
4
6
be_c
Partial dependence
0.54
0.52
0.50
0.48
0.46
0
2000
4000
6000
8000
10000
LP6
Partial dependence
0.55
0.50
0.45
0.40
0.35
0
100
200
300
as_c

Figure SI_2. Partial Dependence Plots for the 10 most influential variables in the BRT5 model. Variables are defined in Table SI_2. Blue hatch lines on the x-axis indicate the deciles of the variable in the model training dataset.



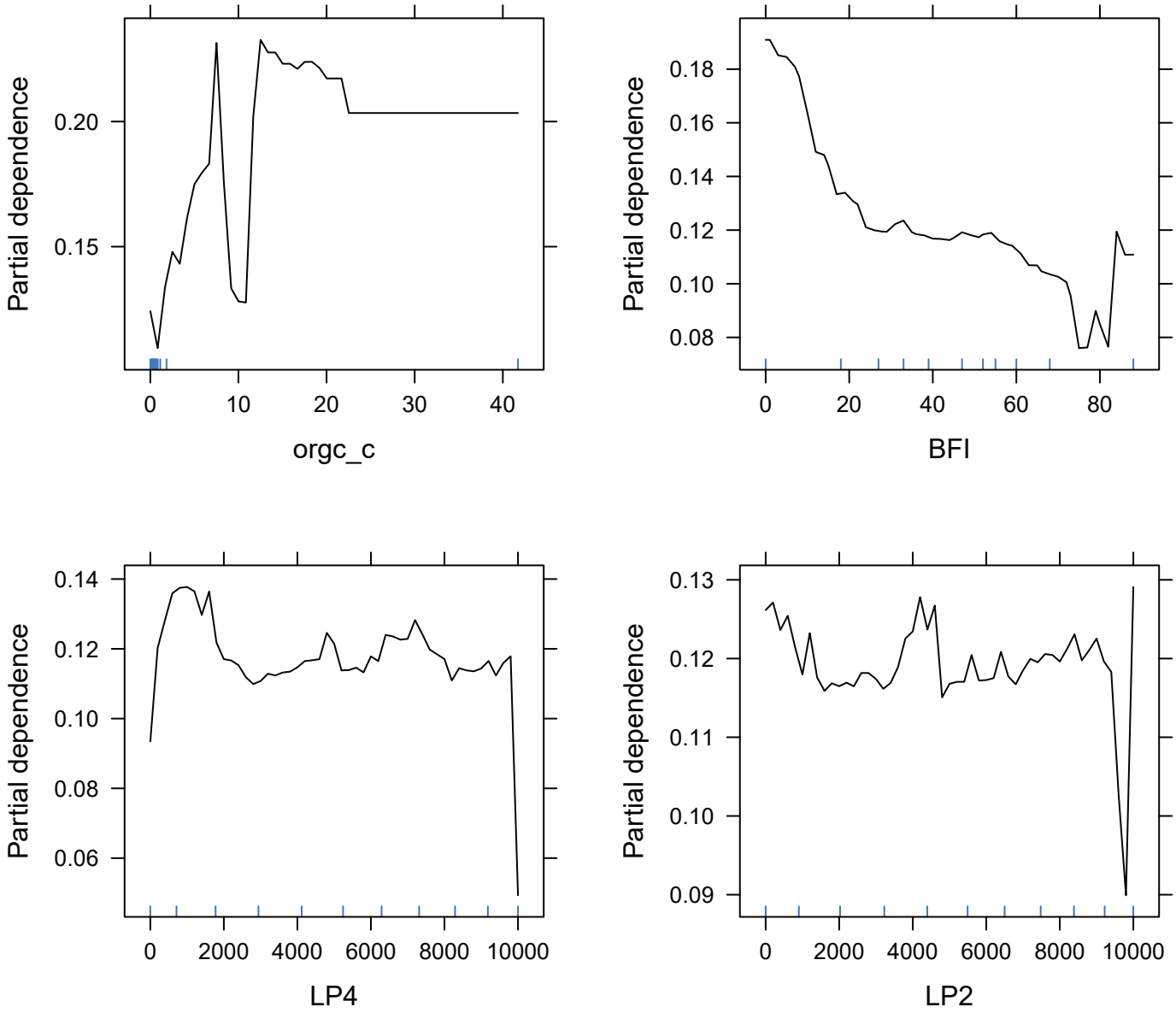
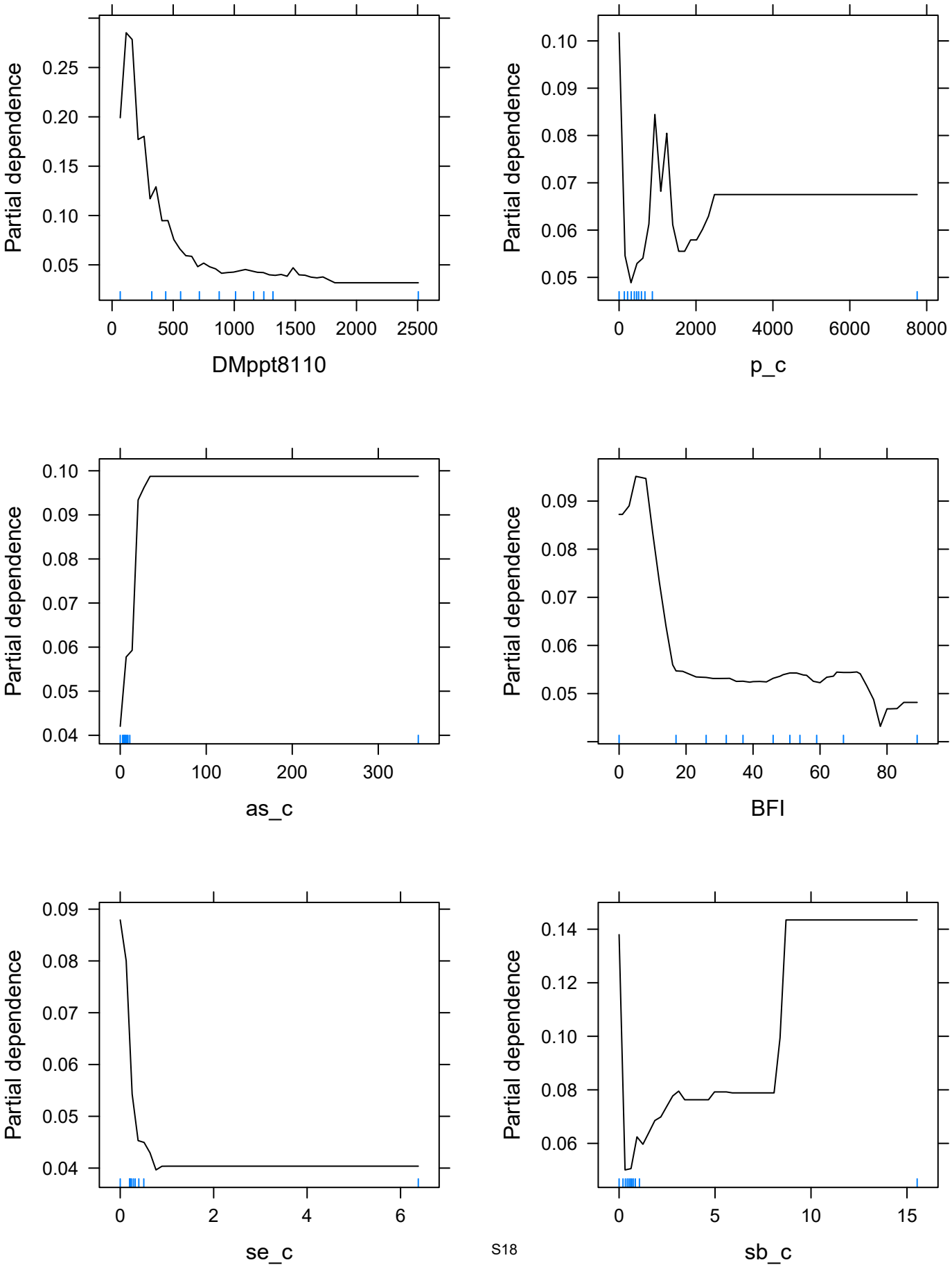

Partial dependence
0.20
0.15
0
10
20
30
40
orgc_c
Partial dependence
0.18
0.16
0.14
0.12
0.10
0.08
0
20
40
60
80
BFI
Partial dependence
0.14
0.12
0.10
0.08
0.06
0
2000
4000
6000
8000
10000
LP4
Partial dependence
0.13
0.12
0.11
0.10
0.09
0
2000
4000
6000
8000
10000
LP2

Figure SI_3. Partial Dependence Plots for the 10 most influential variables in the BRT10 model. Variables are defined in Table SI_2. Blue hatch lines on the x-axis indicate the deciles of the variable in the model training dataset.



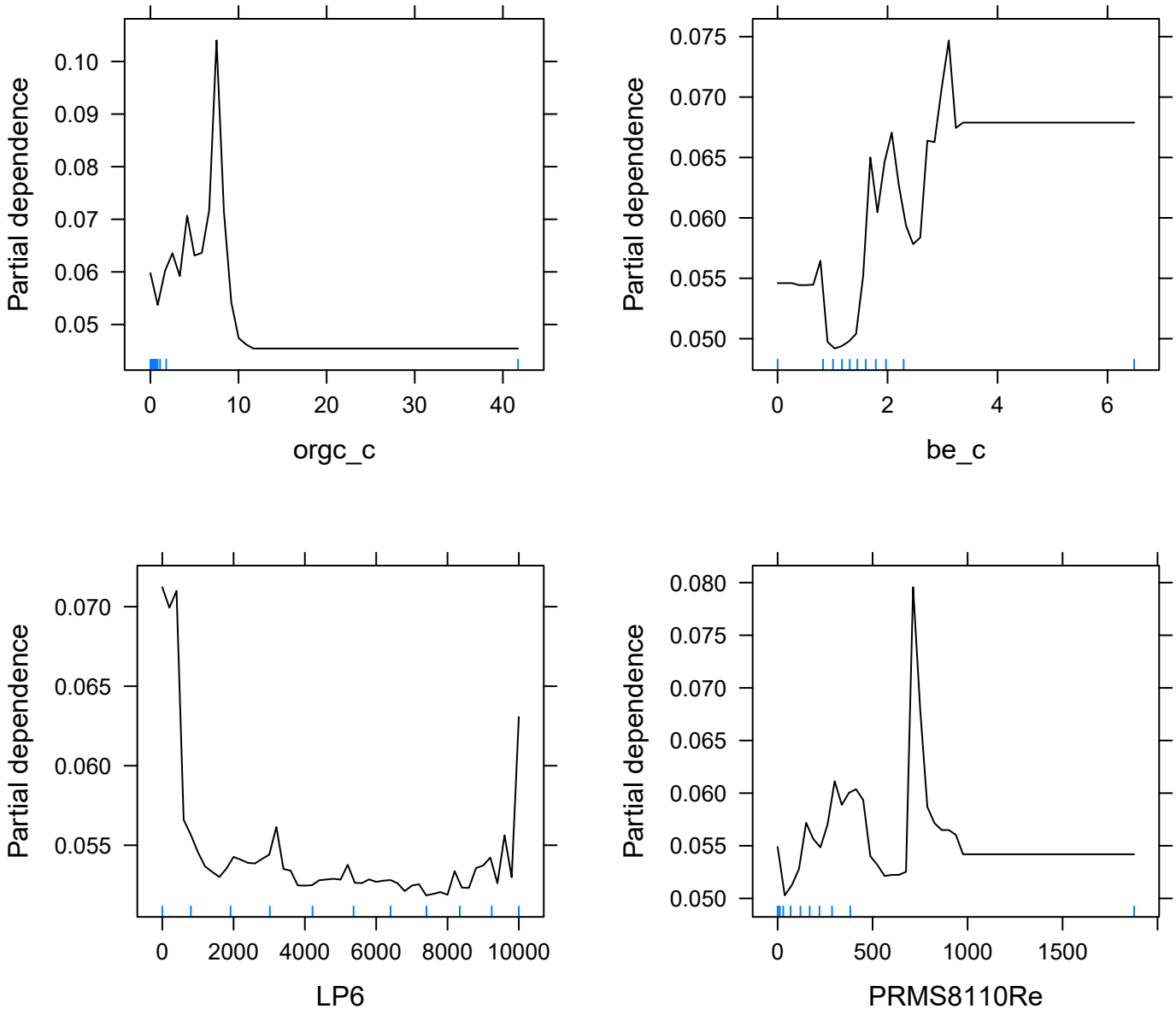
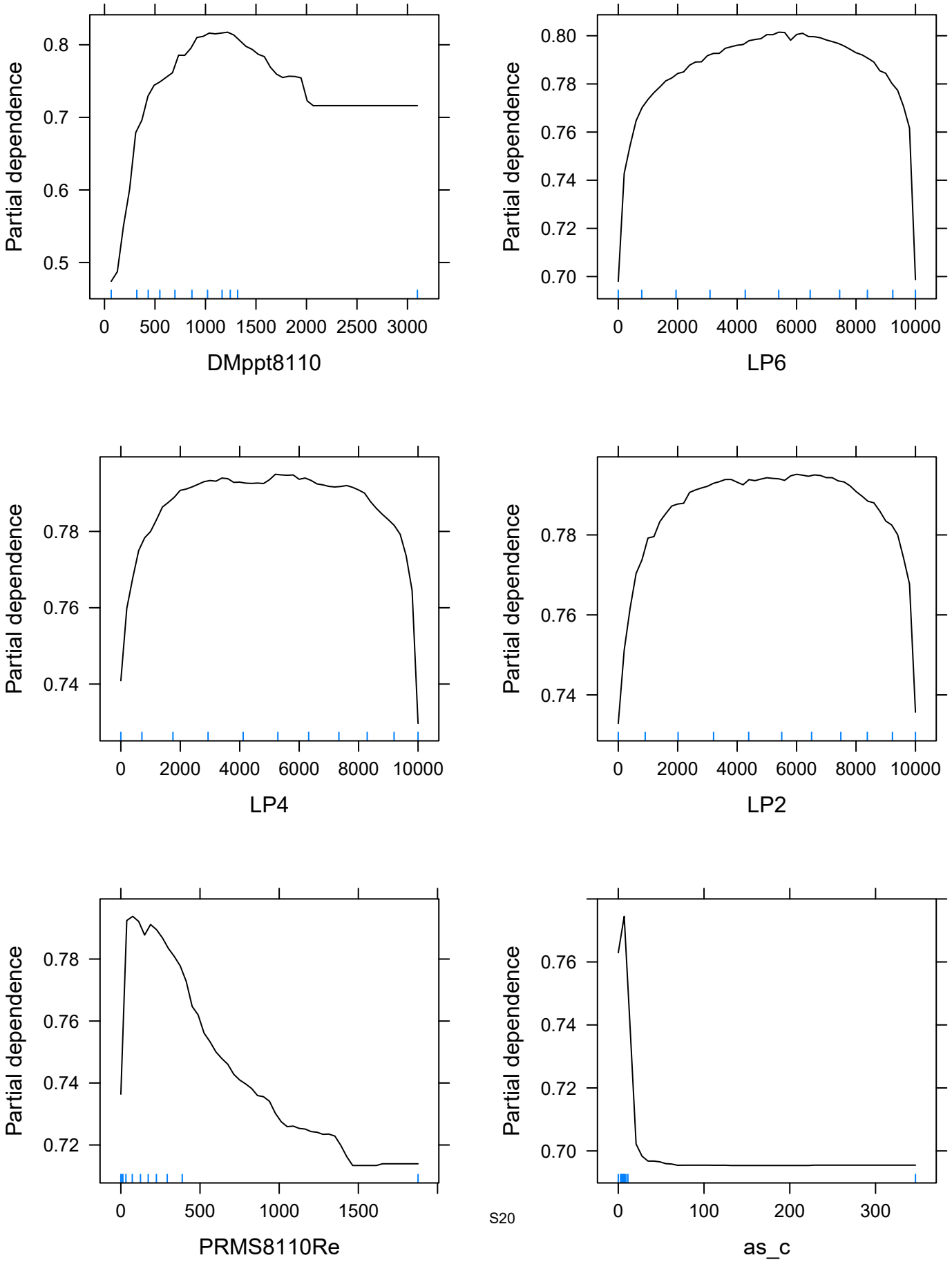

Partial dependence
0.05
0.06
0.07
0.08
0.09
0.10
0
10
20
30
40
orgc_c
Partial dependence
0.050
0.055
0.060
0.065
0.070
0.075
0
2
4
6
be_c
Partial dependence
0.055
0.060
0.065
0.070
0
2000
4000
6000
8000
10000
LP6
Partial dependence
0.050
0.055
0.060
0.065
0.070
0.075
0.080
0
500
1000
1500
PRMS8110Re

Figure SI_4. Partial Dependence Plots for the 10 most influential variables in the RFC model for classification 1, arsenic ≤ 5 µg/L. Variables are defined in Table SI_2. Blue hatch lines on the x-axis indicate the deciles of the variable in the model training dataset.



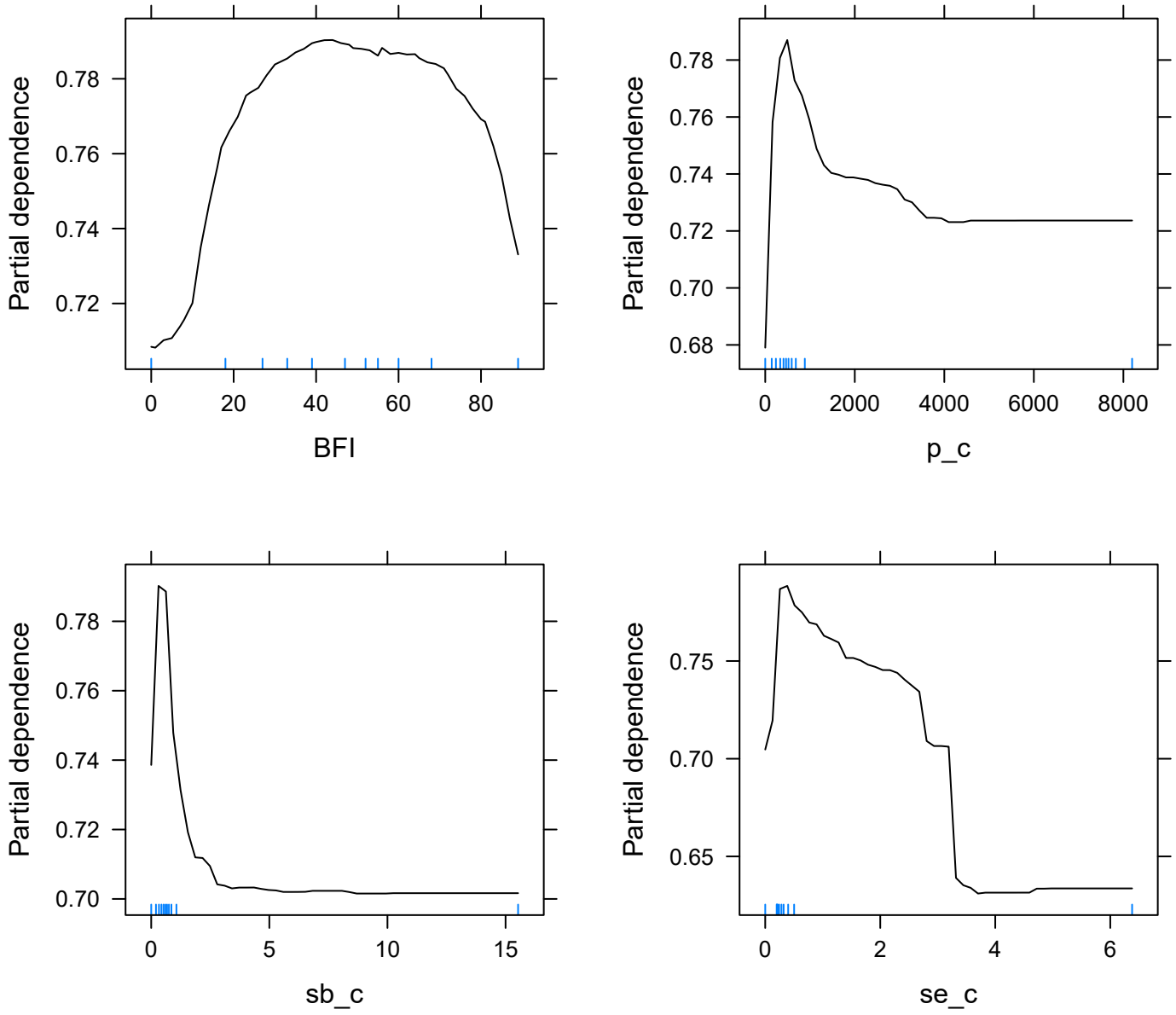
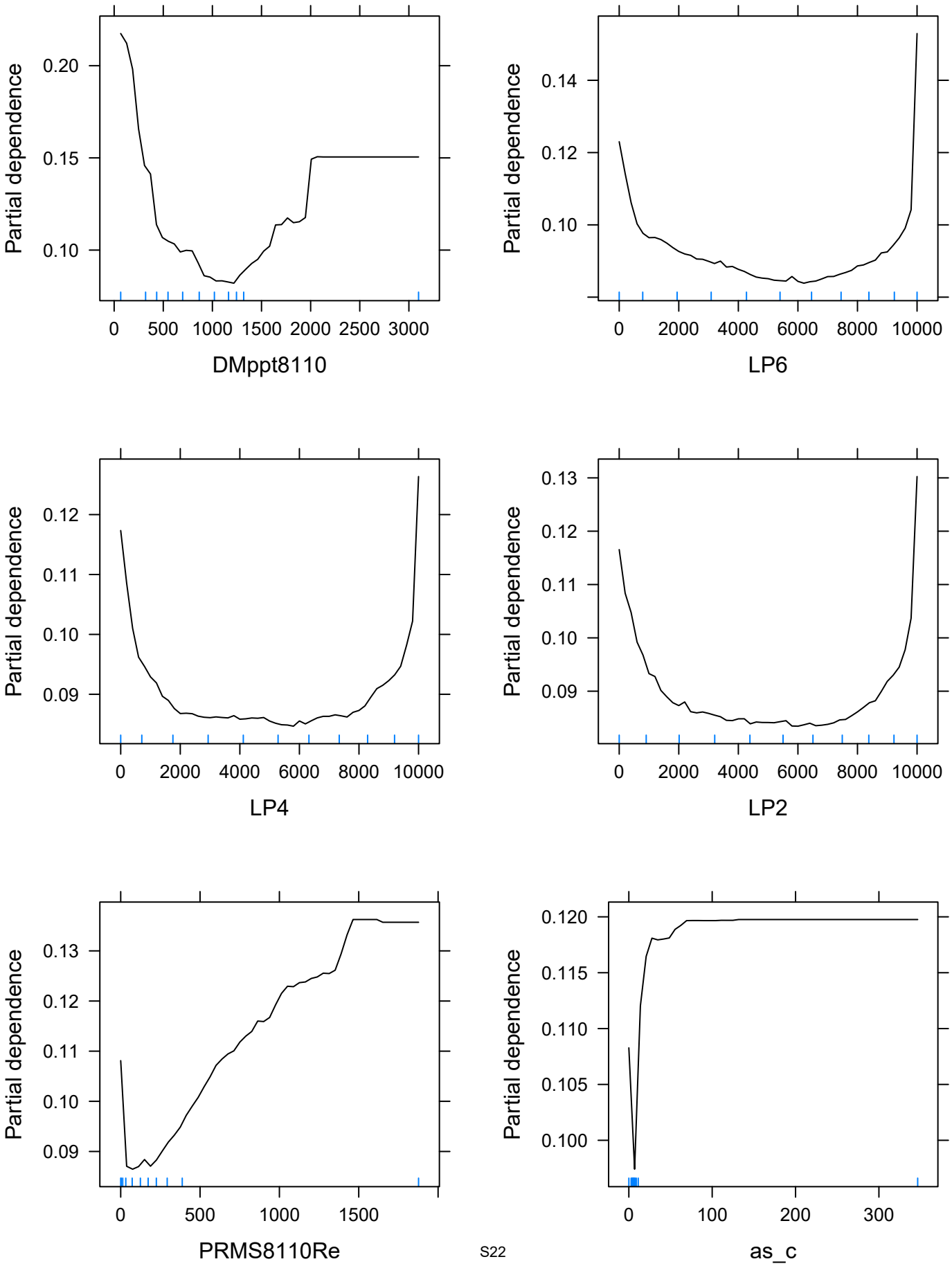

Partial dependence
0.78
0.76
0.74
0.72
0
20
40
60
80
BFI
Partial dependence
0.78
0.76
0.74
0.72
0.70
0.68
0
2000
4000
6000
8000
p_c
Partial dependence
0.78
0.76
0.74
0.72
0.70
0
5
10
15
sb_c
Partial dependence
0.75
0.70
0.65
0
2
4
6
se_c

Figure SI_5. Partial Dependence Plots for the 10 most influential variables in the RFC model for classification 2, arsenic >5 and ≤10 µg/L. Variables are defined in Table SI_2. Blue hatch lines on the x-axis indicate the deciles of the variable in the model training dataset.



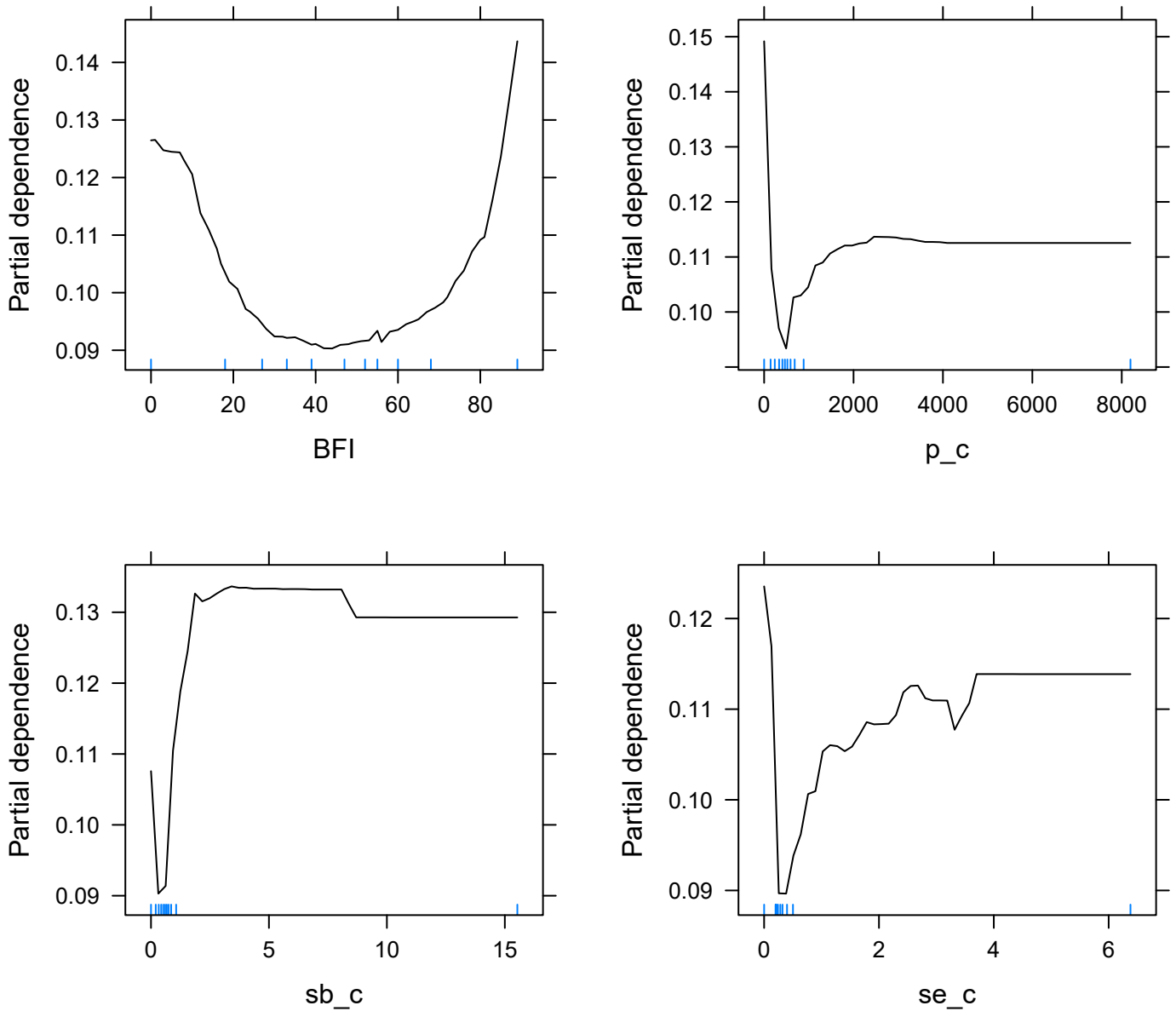
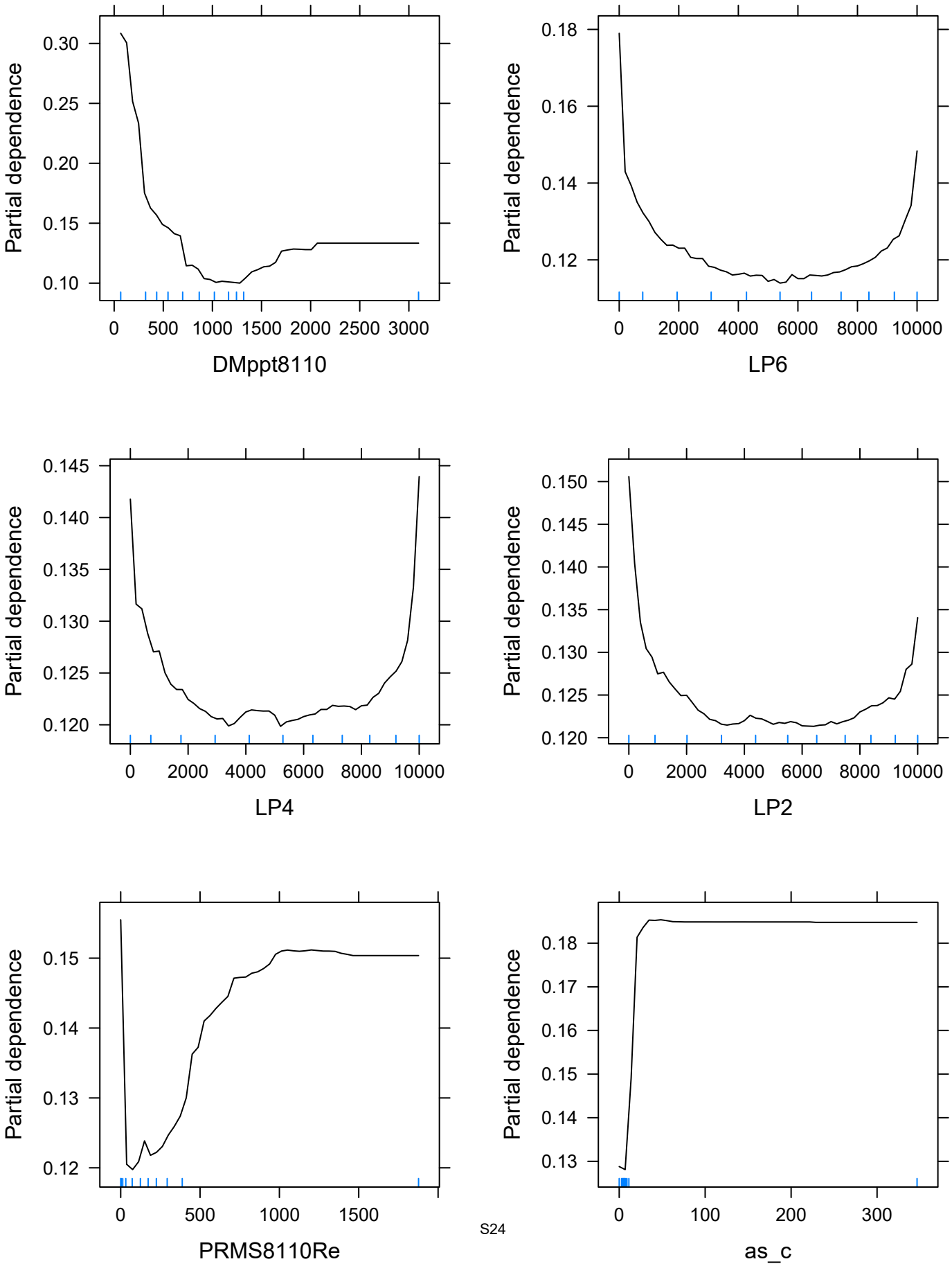
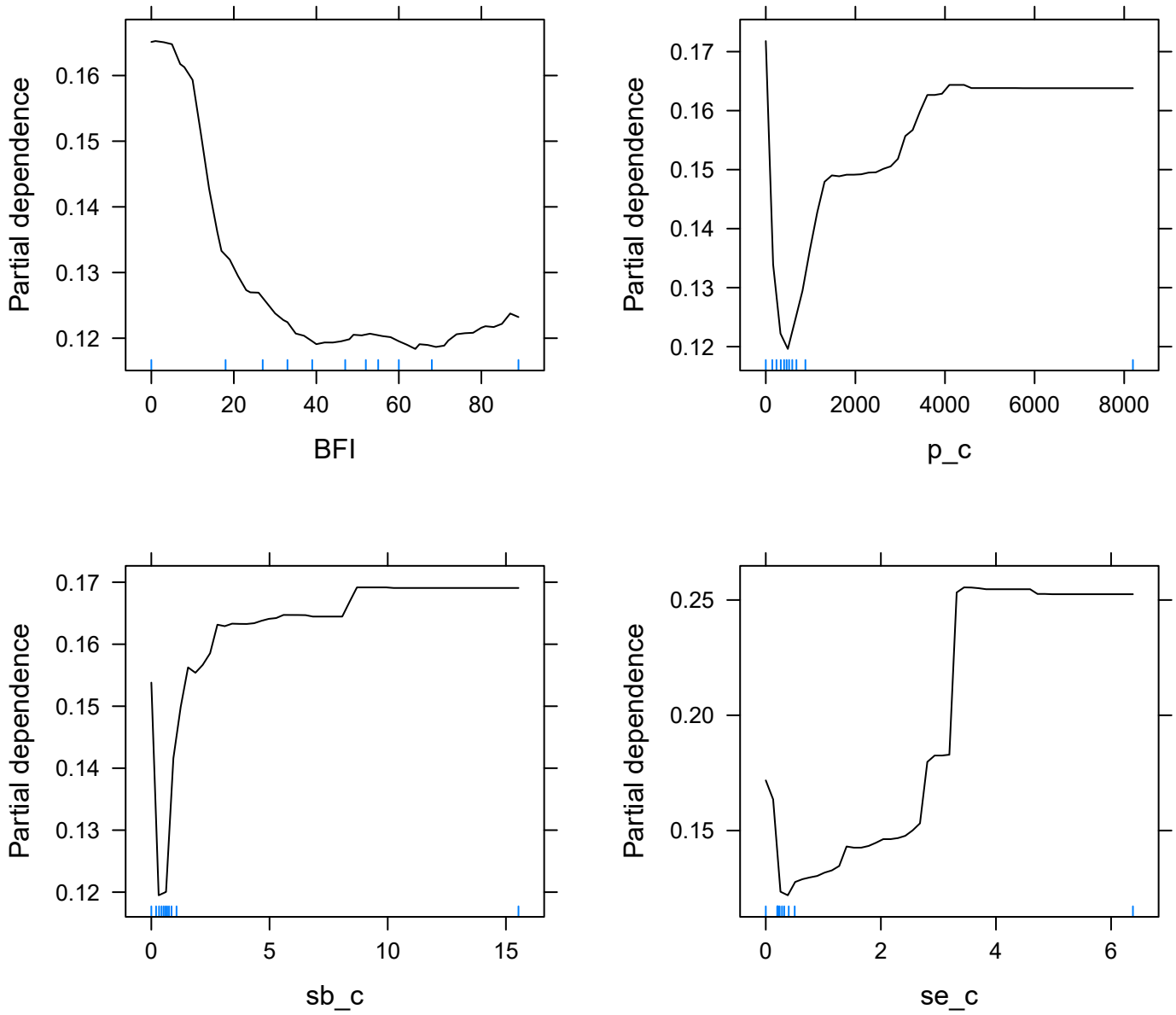

Partial dependence
0.09
0.10
0.11
0.12
0.13
0.14
0
20
40
60
80
BFI
Partial dependence
0.10
0.11
0.12
0.13
0.14
0.15
0
2000
4000
6000
8000
p_c
Partial dependence
0.09
0.10
0.11
0.12
0.13
0
5
10
15
sb_c
Partial dependence
0.09
0.10
0.11
0.12
0
2
4
6
se_c

Figure SI_6. Partial Dependence Plots for the 10 most influential variables in the RFC model for classification 3, arsenic >10 µg/L. Variables are defined in Table SI_2. Blue hatch lines on the x-axis indicate the deciles of the variable in the model training dataset.




Partial dependence
0.12
0.13
0.14
0.15
0.16
0
20
40
60
80
BFI
Partial dependence
0.12
0.13
0.14
0.15
0.16
0.17
0
2000
4000
6000
8000
p_c
Partial dependence
0.12
0.13
0.14
0.15
0.16
0.17
0
5
10
15
sb_c
Partial dependence
0.15
0.20
0.25
0
2
4
6
se_c

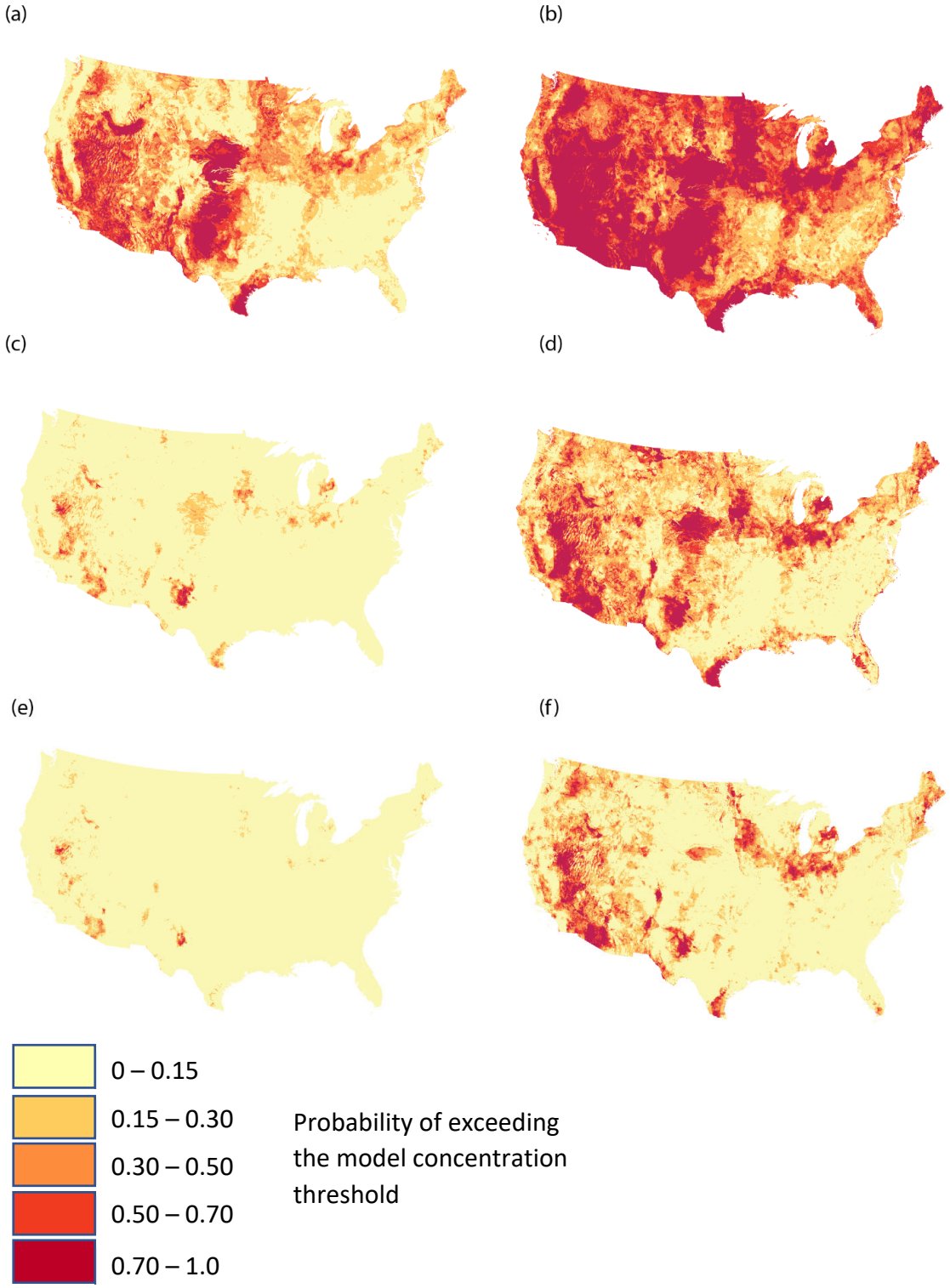



Figure SI_7. Confidence intervals for the model estimates (a) 95% confidence lower bound for the BRT1 model (b) 95% confidence upper bound for the BRT1 model (c) 95% confidence lower bound for the BRT5 model (d) 95% confidence upper bound for the BRT5 model (e) 95% confidence lower bound for the BRT10 model (f) 95% confidence upper bound for the BRT10 model. Confidence intervals were calculated following the methods developed by Ransom et al., 2017, A hybrid machine learning model to predict and visualize nitrate concentration throughout the Central Valley aquifer, California, USA, *Sci Tot Env,*601-602:1160-1172, http://dx.doi.org/10.1016/j.scitotenv.2017.05.192.