Supporting Information for

Accurate and precise prediction of soil properties from a large mid infrared spectral library

Shree R. S. Dangal[1,†,*] and Jonathan Sanderman[1,†,*], Skye Wills[2], Leonardo Ramires-Lopez[3]

[1]The Woods Hole Research Center, 149 Woods Hole Road, Falmouth, MA-02540, USA; sdangal@whrc.org (SRSD); jsanderman@whrc.org (JS)

[2]United States Department of Agriculture – Natural Resources Conservation Service (USDA-NRCS), 100 Centennial Mall North, Lincoln, NE-68508, USA; skye.wills@lin.usda.gov

[3]BUCHI Labortechnik AG, Meierseggstrasse 40, 9230 Flawil, Switzerland; ramirez.lopez.leo@gmail.com

*Correspondence: sdangal@whrc.org (SRSD); jsanderman@whrc.org (JS)

[†]These authors contributed equally to this work.

Contents of this file

Table S1. Performance of PLSR, MBL, RF and Cubist models on independent validation sets for non-normally distributed OC and near normally distributed pH, using square root, box-cox and log transformation, and no transformation.

Soil		Square root			Box-cox			Log			Untransformed		
		Bias	R^2	RMSE	Bias	R^2	RMSE	Bias	R^2	RMSE	Bias	R^2	RMSE
OC	PLSR	0.12	0.99	1.19	0.04	0.97	1.92	0.05	0.98	1.83	-0.12	0.98	1.55
(N=8498)	MBL	0.03	0.99	0.64	0.07	0.99	0.70	0.06	0.99	0.7	-0.07	0.99	0.68
	RF	0.11	0.99	0.93	0.14	0.99	0.97	0.14	0.99	0.97	0.04	0.99	0.93
	Cubist	0.01	0.99	0.69	0.01	0.99	0.64	0.01	0.99	0.69	0.03	0.99	0.68
pH	PLSR	0.04	0.74	0.54	0.04	0.74	0.54	0.06	0.73	0.54	0.02	0.74	0.54
(N=6990)	MBL	0	0.89	0.34	-0.01	0.89	0.34	0	0.89	0.34	-0.01	0.89	0.34
	RF	0	0.82	0.45	0	0.82	0.45	0.01	0.82	0.45	-0.01	0.82	0.45
	Cubist	0.01	0.88	0.36	0	0.89	0.35	0.01	0.85	0.4	0.01	0.88	0.36

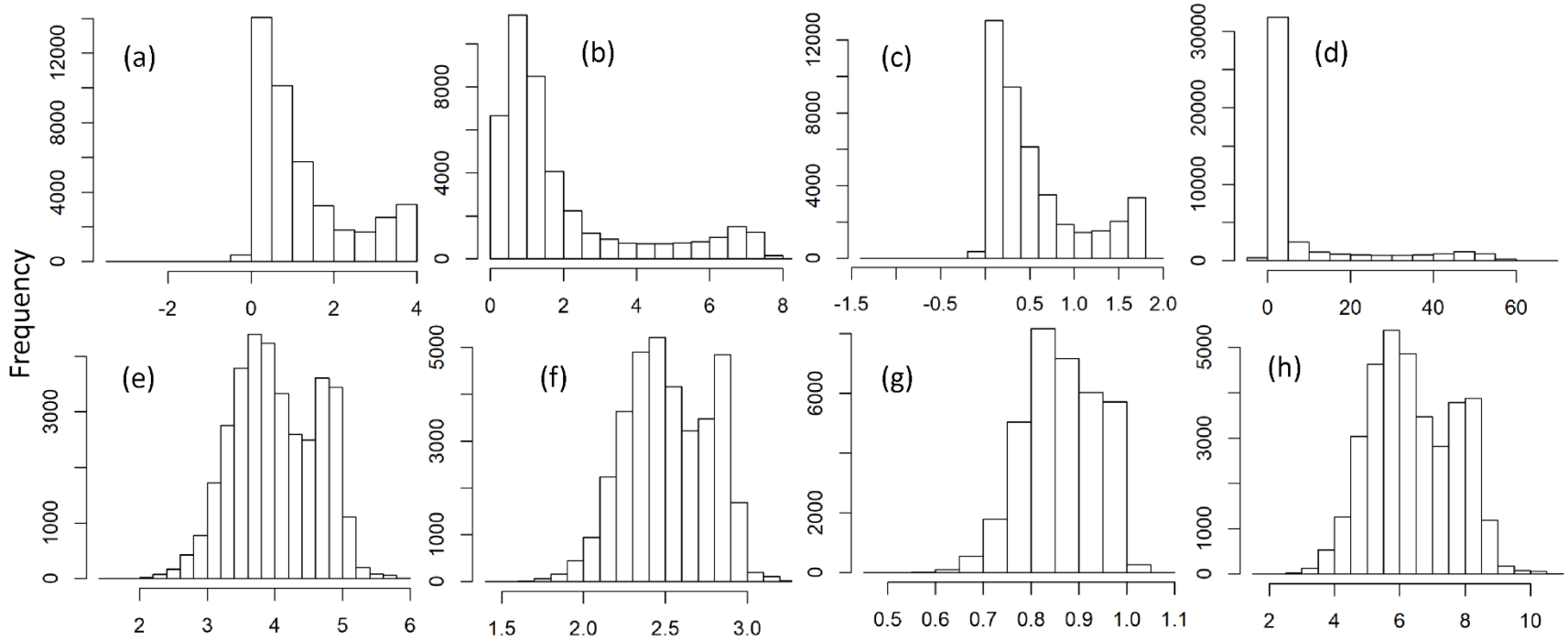

Figure S1. Histogram plot of box cox transformed OC (a) and pH (e), square root transformed OC (b) and pH (f), log transformed OC (c) and pH (g) and untransformed OC (d) and pH (h).

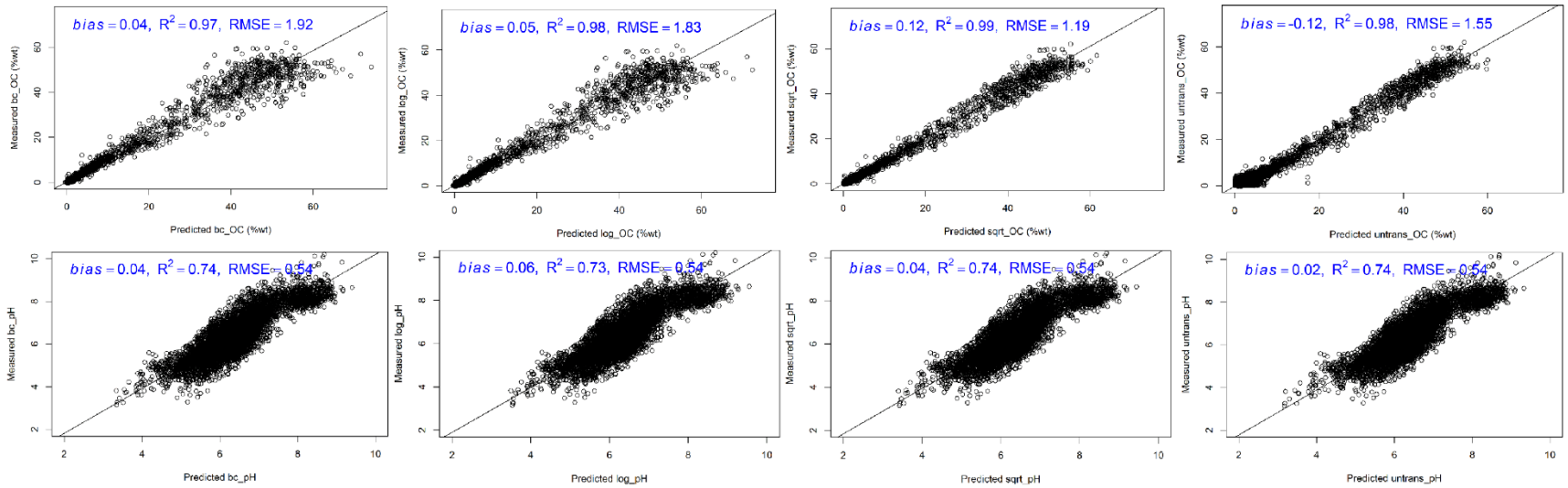

Figure S2. Prediction of transformed and untransformed OC (N = 8498) and pH (N = 6990) using a global PLSR model. All transformed analytical data were back transformed to its original form before assessing the model performance.

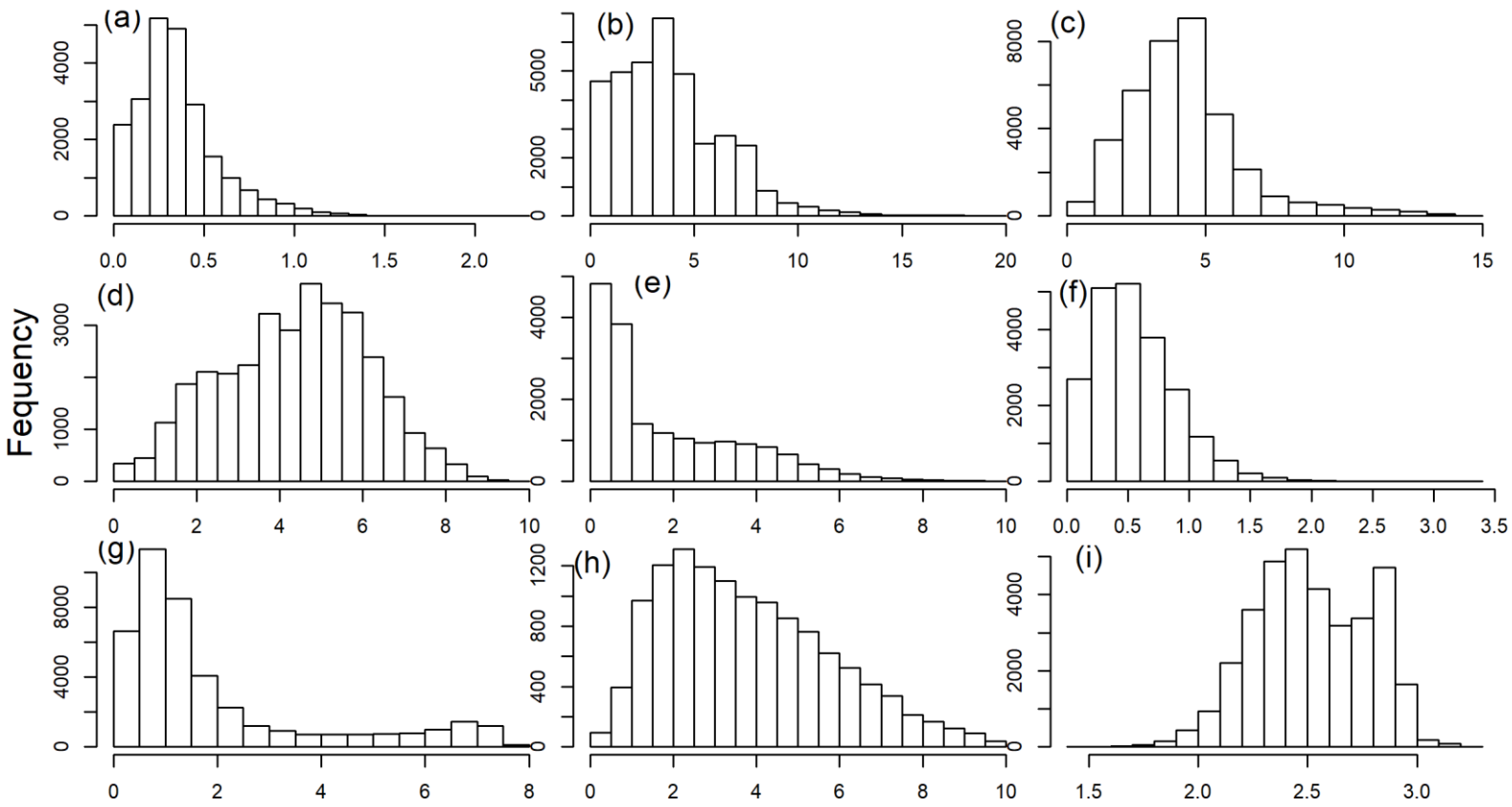

Figure S3. Histogram of square root transformed Al (a), Ca (b), CEC (c), Clay (d), CO_3 (e), Fe (f), OC (g), OCD (h) and pH (i).

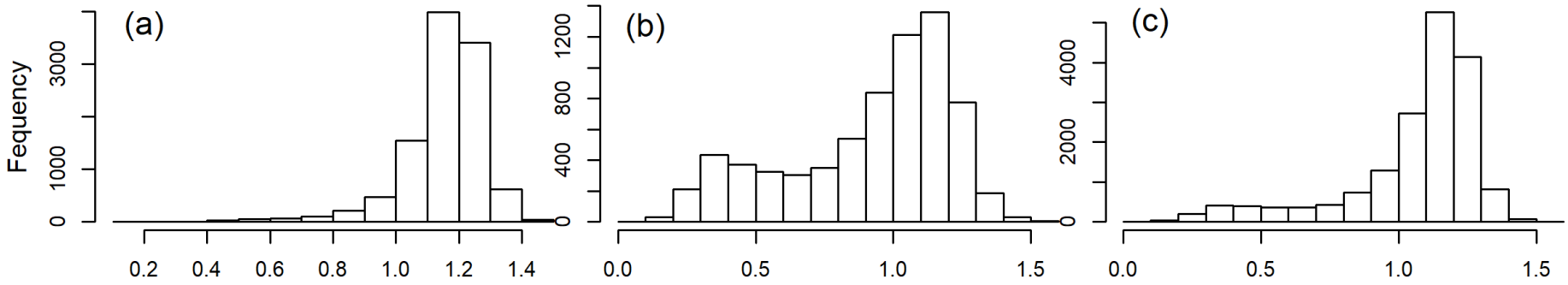

Figure S4. Histogram of square root transformed BD using clod (a), core (b) and combined clod and core (c) methods.

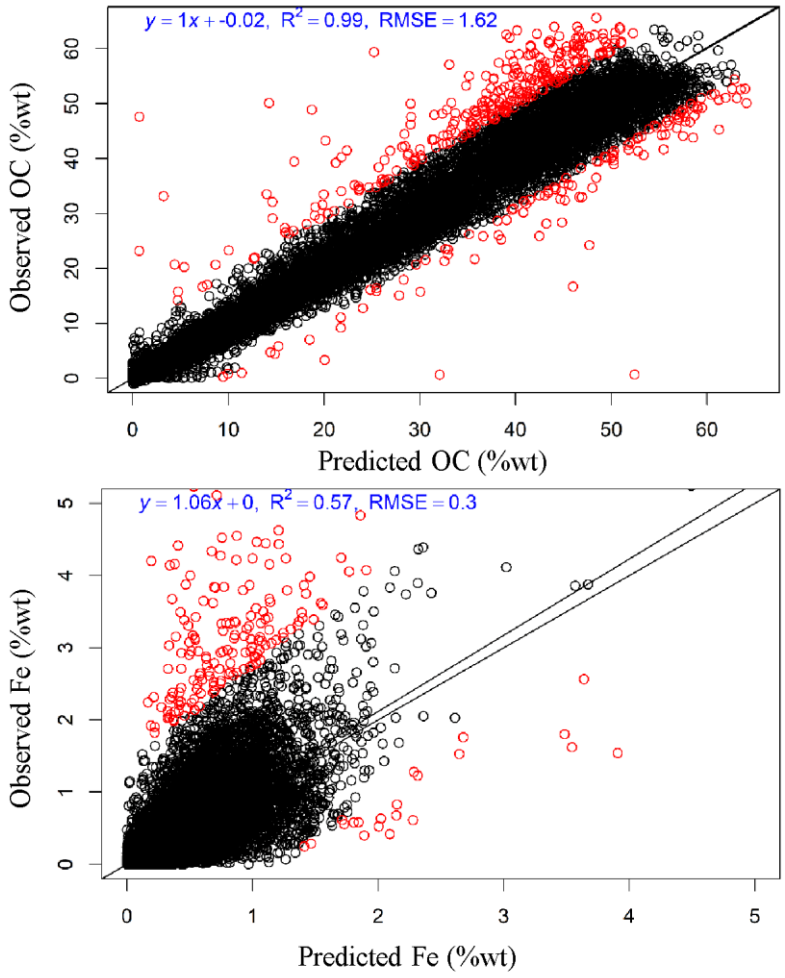

Figure S5. Outlier detection using a PLSR model for OC and Fe. The two soil properties are picked using the best and worst model fit (R2) among different soil properties. The red dots are the samples that have been removed as an outlier, while the black dots are the samples that are retained for developing multivariate regression and machine learning models.

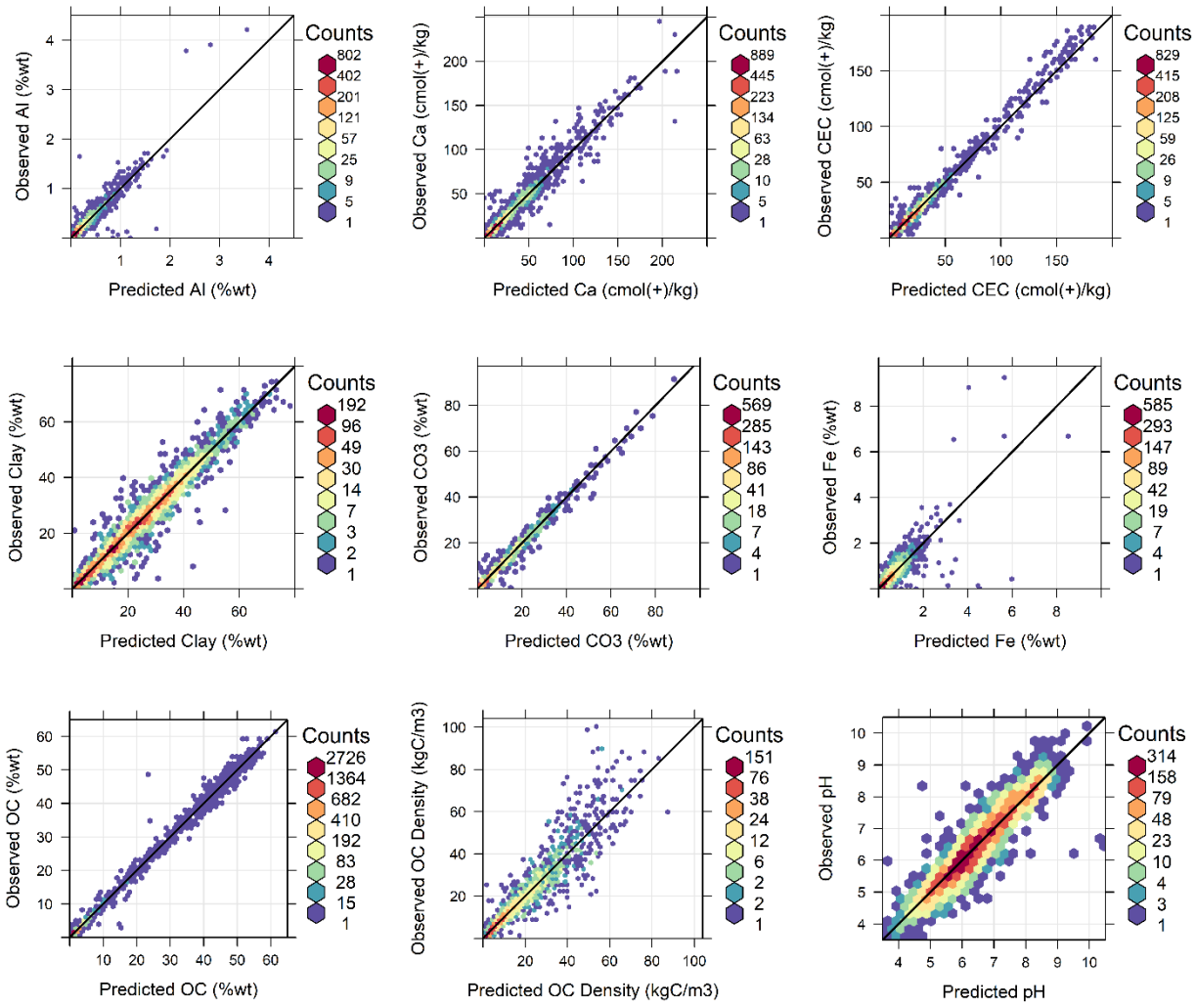

Figure S6. Comparison of laboratory measured soil properties against predicted values using Cubist for an independent validation set. Data were back-transformed before plotting and calculating regression statistics.

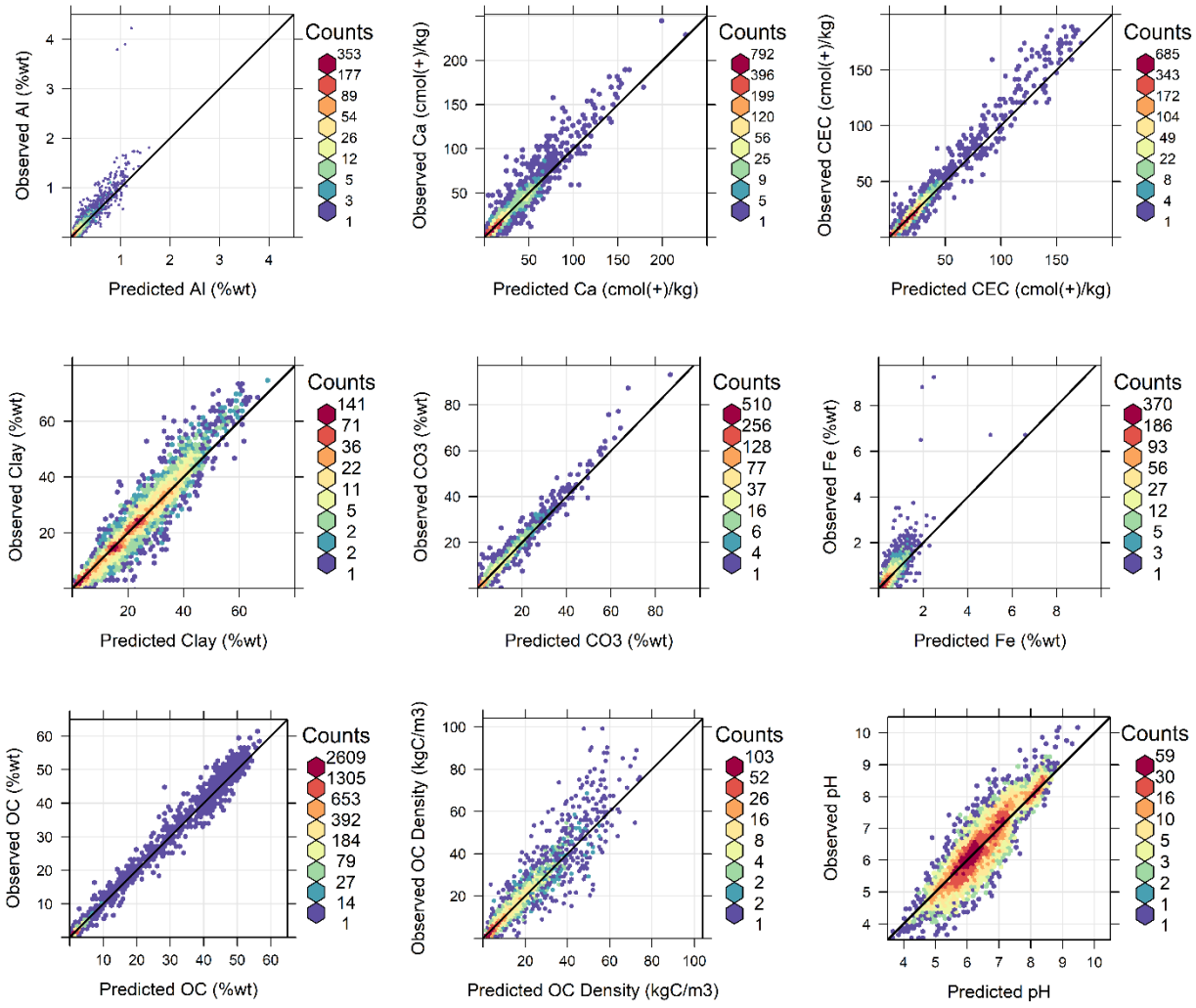

Figure S7. Comparison of laboratory measured soil properties against predicted values using random forest model for an independent validation set. Data were back-transformed before plotting and calculating regression statistics.

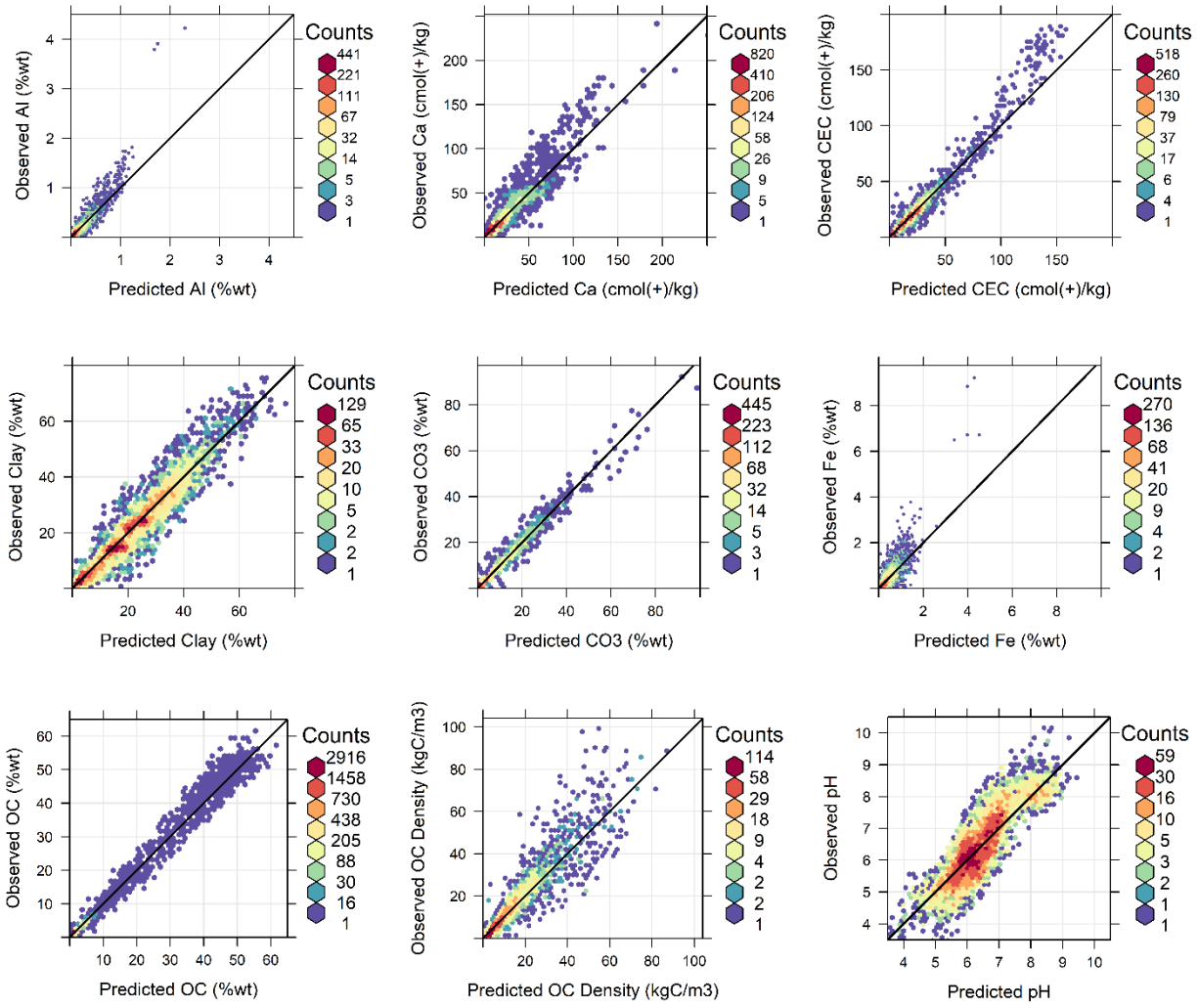

Figure S8. Comparison of laboratory measured soil properties against predicted values using global PLSR for an independent validation set. Data were back-transformed before plotting and calculating regression statistics.

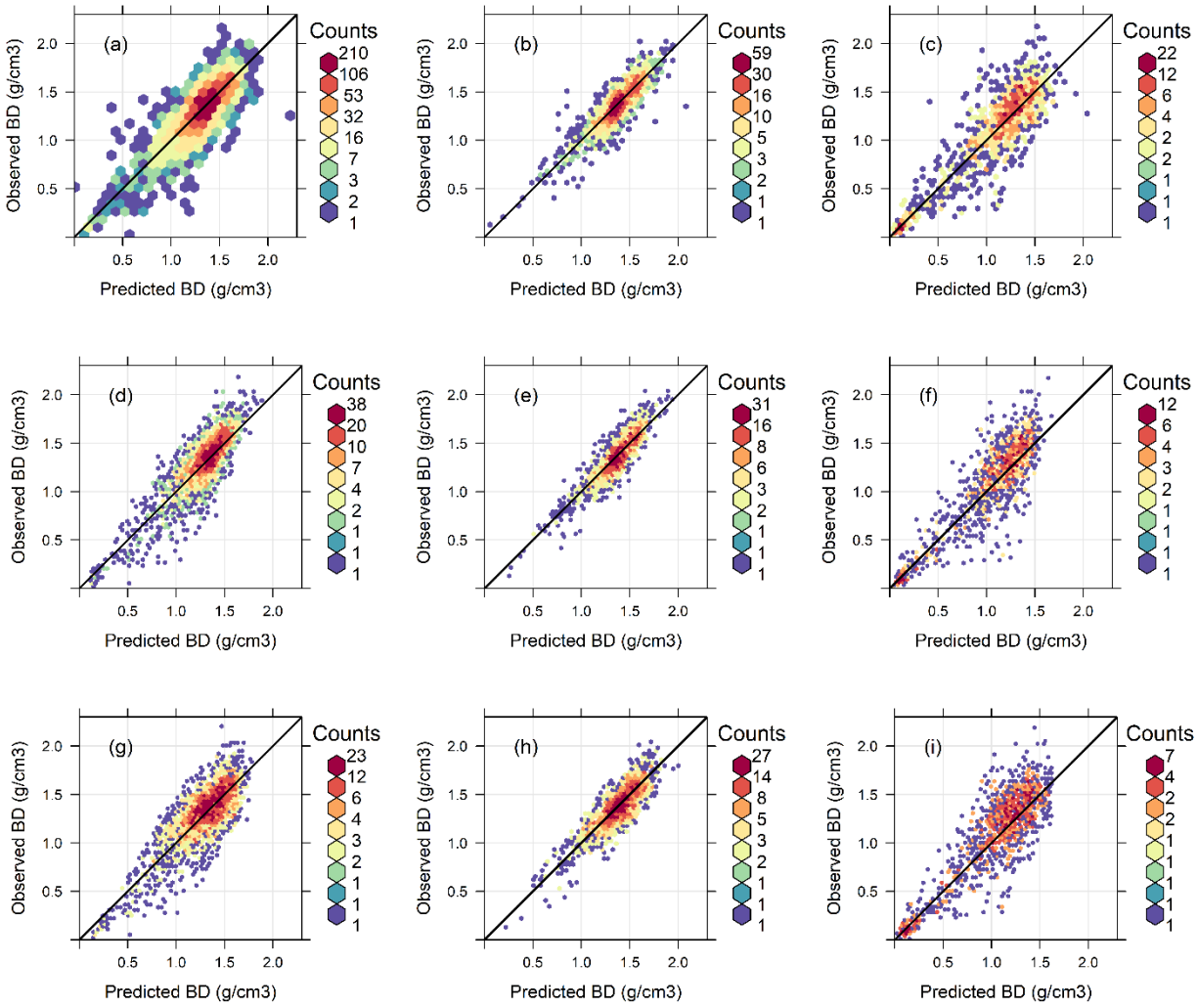

Figure S9. Comparison of bulk density against different analytical methods using cubist -- combined clod and core (a), clod (b) and core (c); random forest -- combined clod and core (d), clod (e) and core (f); and PLSR -- combined clod and core (g), clod (h) and core (i) models.

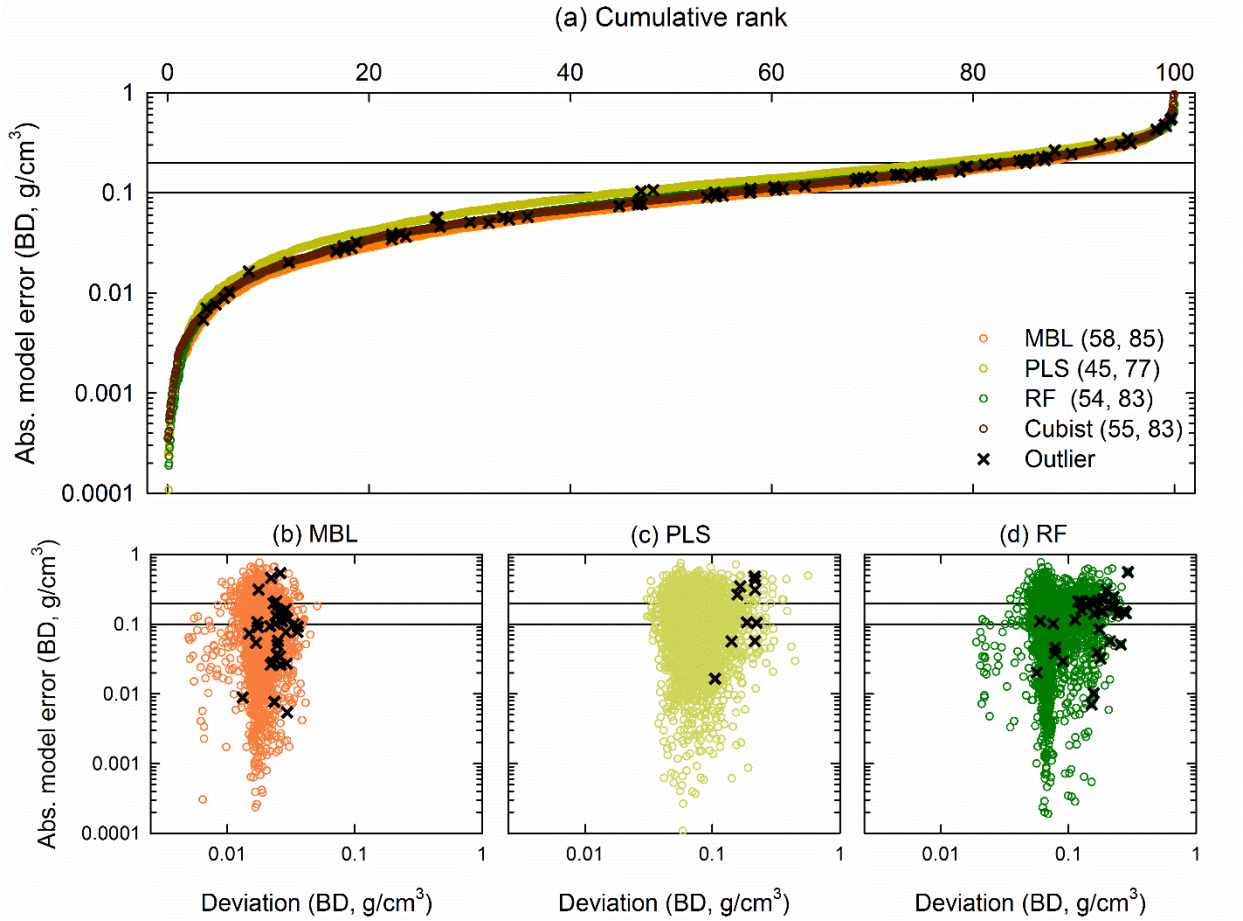

Figure S10. Absolute model error and uncertainty estimates (deviation) of independent validation sets for BD using MBL, PLSR, RF and Cubist models. The top panel figure (A) shows the cumulative rank of the absolute difference between the predicted and observed values (N = 3306). The number in parenthesis are the % of samples above 0.1 and 0.2 g/cm^3 of absolute error. Only absolute error is given for Cubist. The bottom panel figure (b, c & d) shows the relationship between absolute model error and deviation using MBL, PLSR and RF models. The black cross symbols are the samples that were flagged as untrustworthy prediction using MBL, PLSR and RF models.

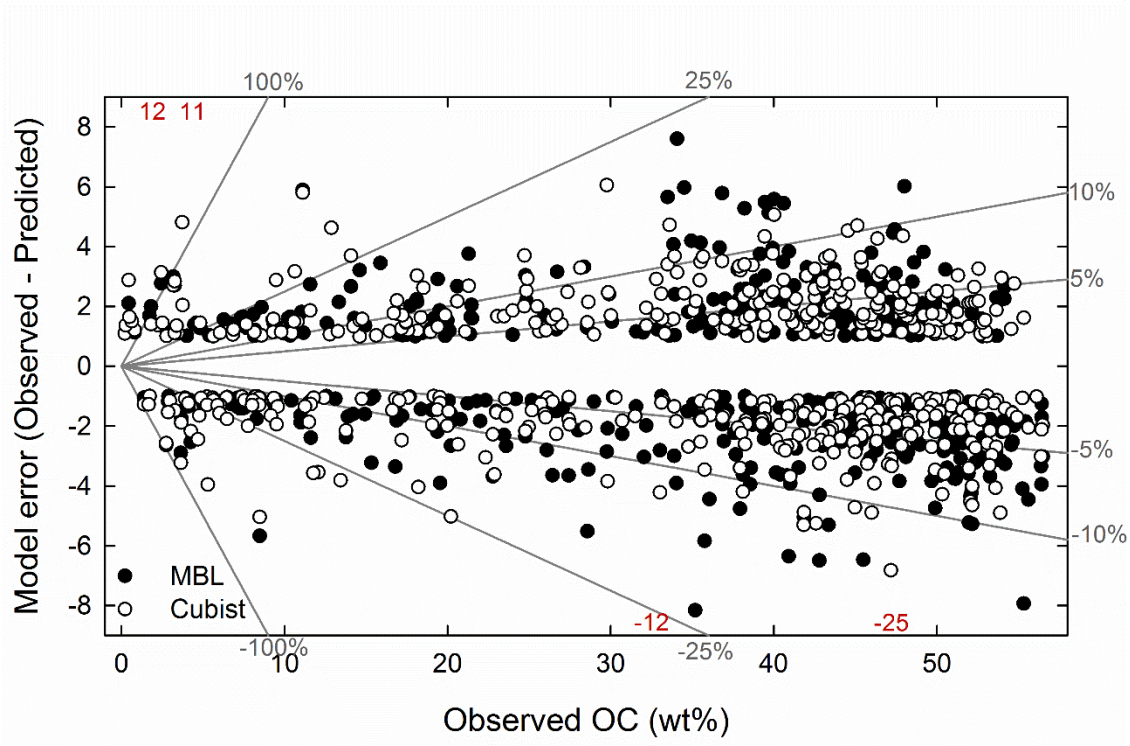

Figure S11. Model error (observed – predicted values) for the poorest OC predictions using Memory-based learner (MBL) and Cubist. Only values with an absolute error > 1.0 wt% are shown. The four red numbers are extreme Cubist model error values where the data point is not shown. Dark grey reference lines indicate the percent relative error (100 x model error / observed value).

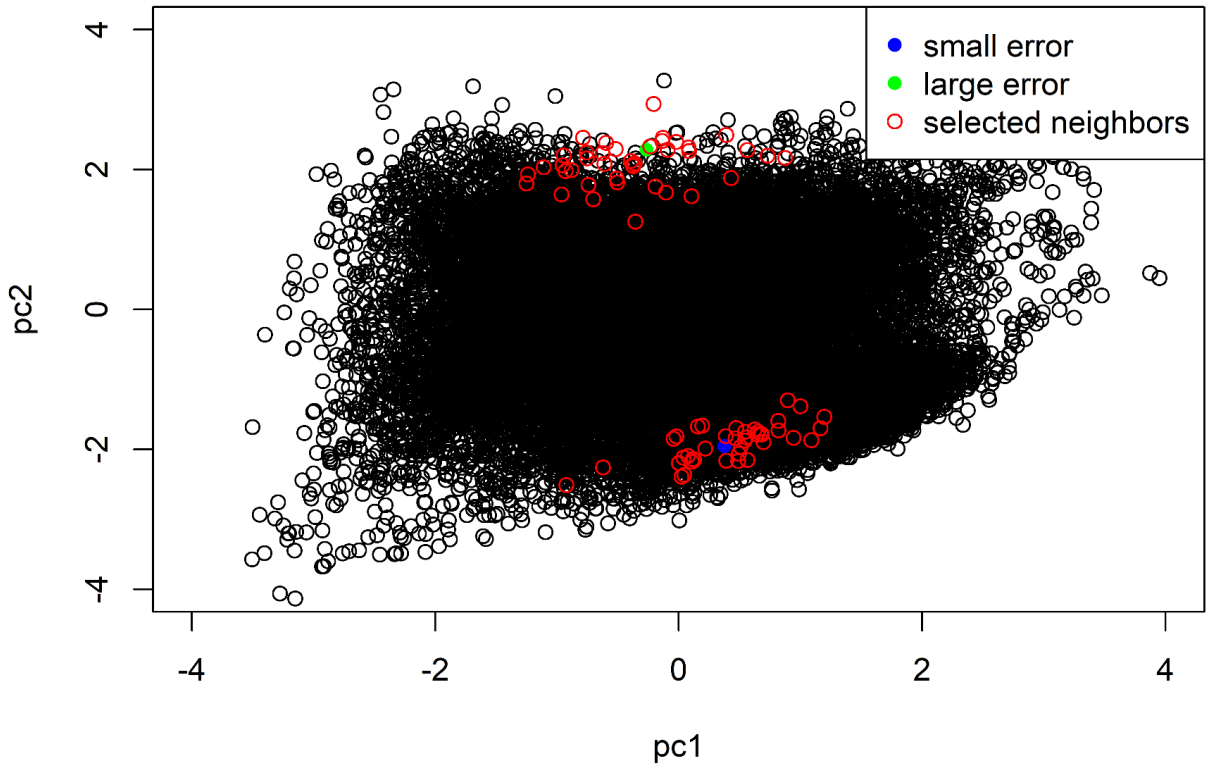

Figure S12. Neighbor selection using the memory-based learner. The green dot represent the sample with the largest prediction error, while the blue dot represent the sample with the smallest prediction error in validation sets. The red dots are the samples used to build the local spectroscopic models.

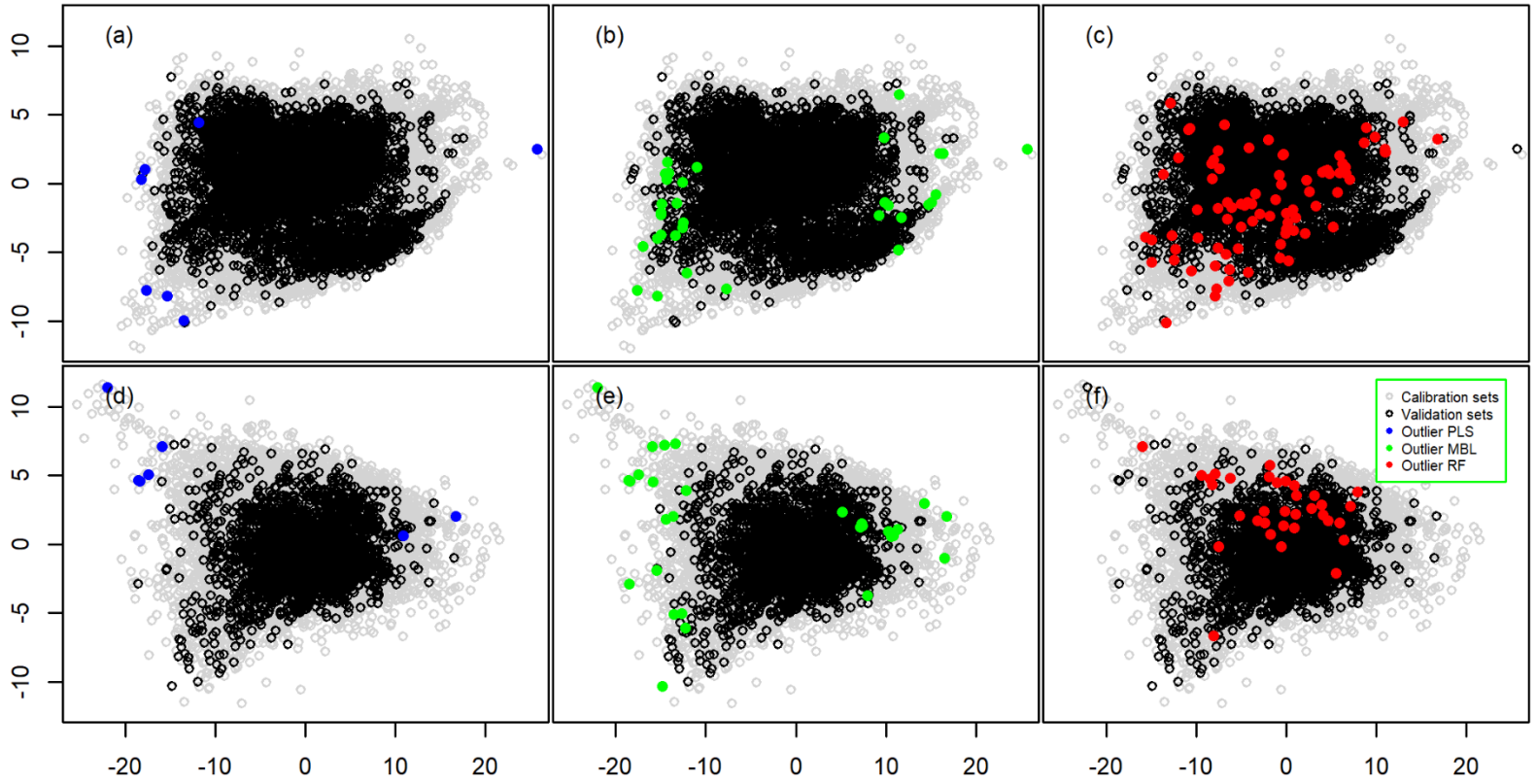

Figure S13. Samples flagged as untrustworthy for OC predicted using PLSR (a), MBL (b) and RF (c) models, and for BD predicted using PLSR (d), MBL (e) and RF (f). The x-axis are the values corresponding to first principal components, while the y-axis are the values corresponding to second principal components.