Supplementary material

João P. Darela-Filho, Anja Rammig, Katrin Fleischer, Tatiana Reichert, Laynara F. Lugli, Carlos Alberto Quesada, Luis Carlos Colocho Hurtarte, Matheus Dantas de Paula, David M. Lapola

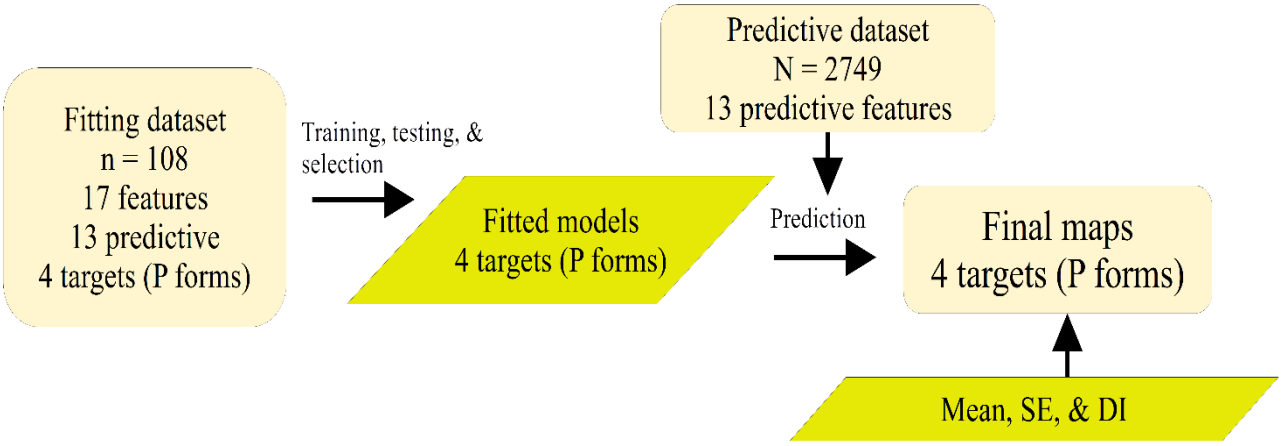


Figure S1: Step flow of operations to generate the estimated maps of P forms in the topsoil profile (0 - 30 cm) of the study area. The fitting dataset (See section 2.1 in the main text) is used to train and test the random forest regression models. Predictive features and P forms (total, available, organic, and inorganic P) are listed in table 1. In this stage the models are selected based on accuracy and cross validation scores (Table 3, main text). The predictive dataset (See section 2.2 in the main text) is used to predict the P forms using the selected random forest regression models. In the final stage the predicted maps for each P form are aggregated using the mean and the standard error (SE) is calculated. Some of the grid cells in the final maps are excluded based on a dissimilarity index (DI) estimated using the fitting and the predictive datasets.

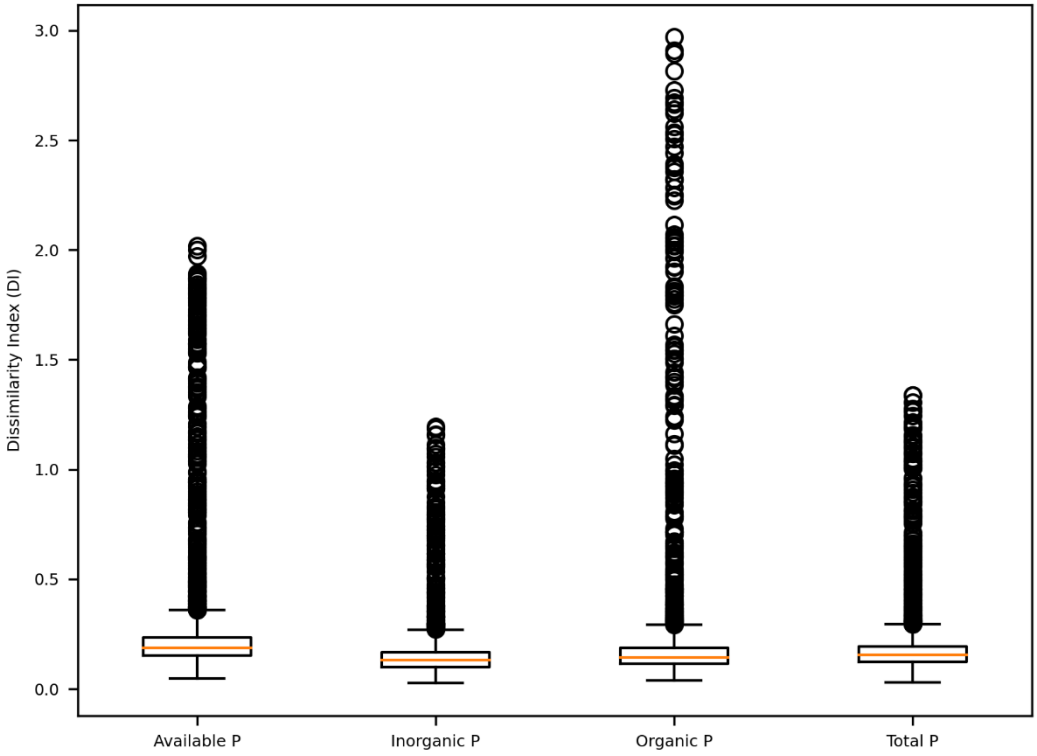


Figure S2: Distribution of the Dissimilarity Index for 2749 grid cells of the predictive dataset, calculated based on Meyer and Pebesma (2021). The outliers in the upper bound are the grid cells of the predictive dataset that are excluded (masked) from the prediction of the final maps in the figures of the main text.

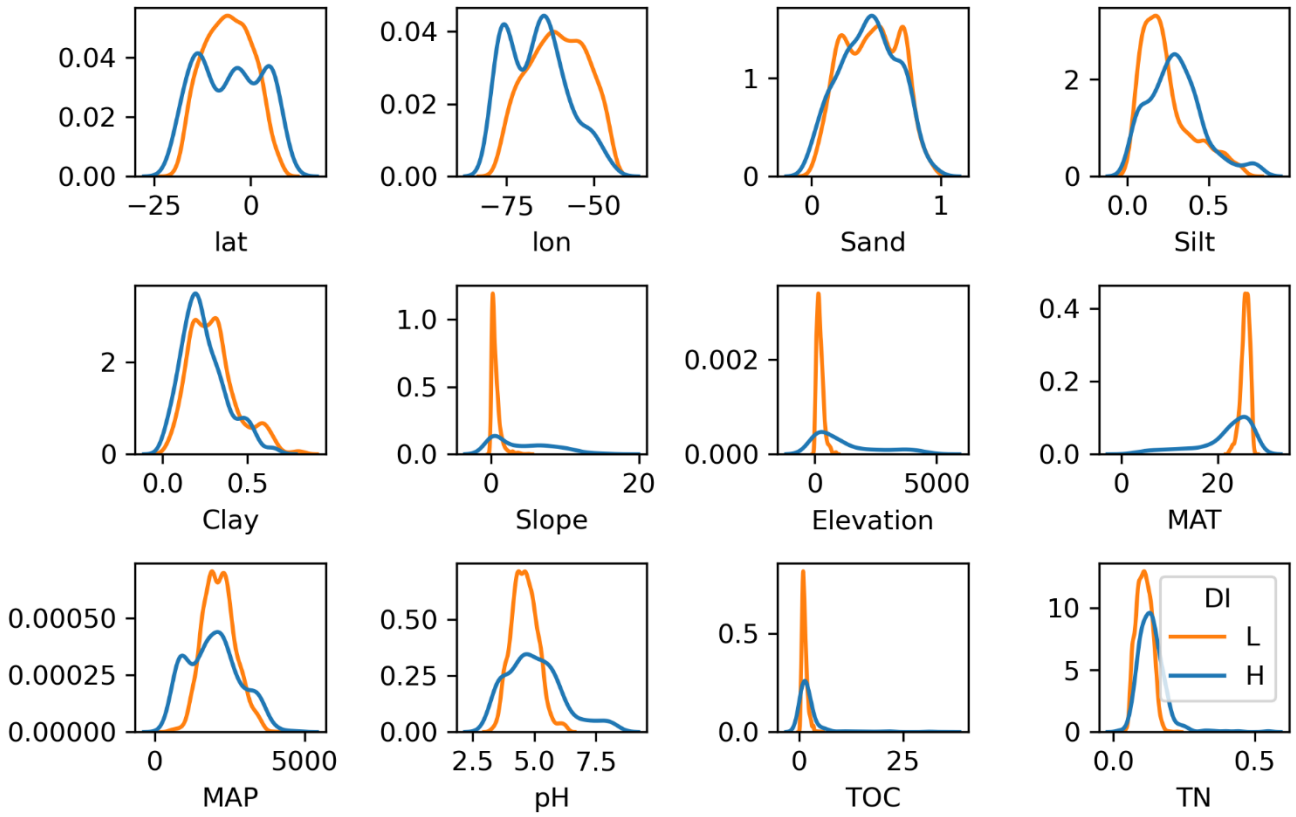


Figure S3: Kernel density plots of the features in the predictive dataset based on the values of the Dissimilarity Index (DI). The group of excluded grid cells presented values higher (H, in blue) than the defined DI threshold while the non-excluded grid cells presented values lower (L, in orange) than the DI thresholds defined. The excluded values are based on the intersection of the masked values for each P form and represents 16.35 % of the area covered by the predictive dataset.

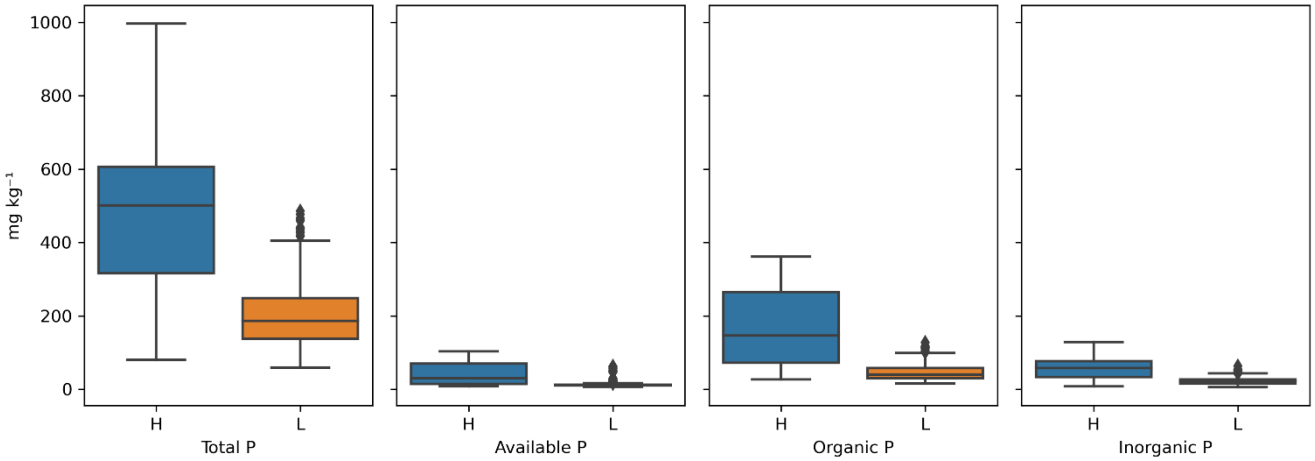


Figure S4: Predicted values of P forms according to the exclusion based on the dissimilarity index (DI). The group of excluded grid cells presented values higher (H) than the defined DI threshold while the non-excluded grid cells presented values lower (L) than the DI thresholds defined. The excluded values are based on the intersection of the masked values for each P form and represents 16.35 % of the area covered by the predictive dataset.

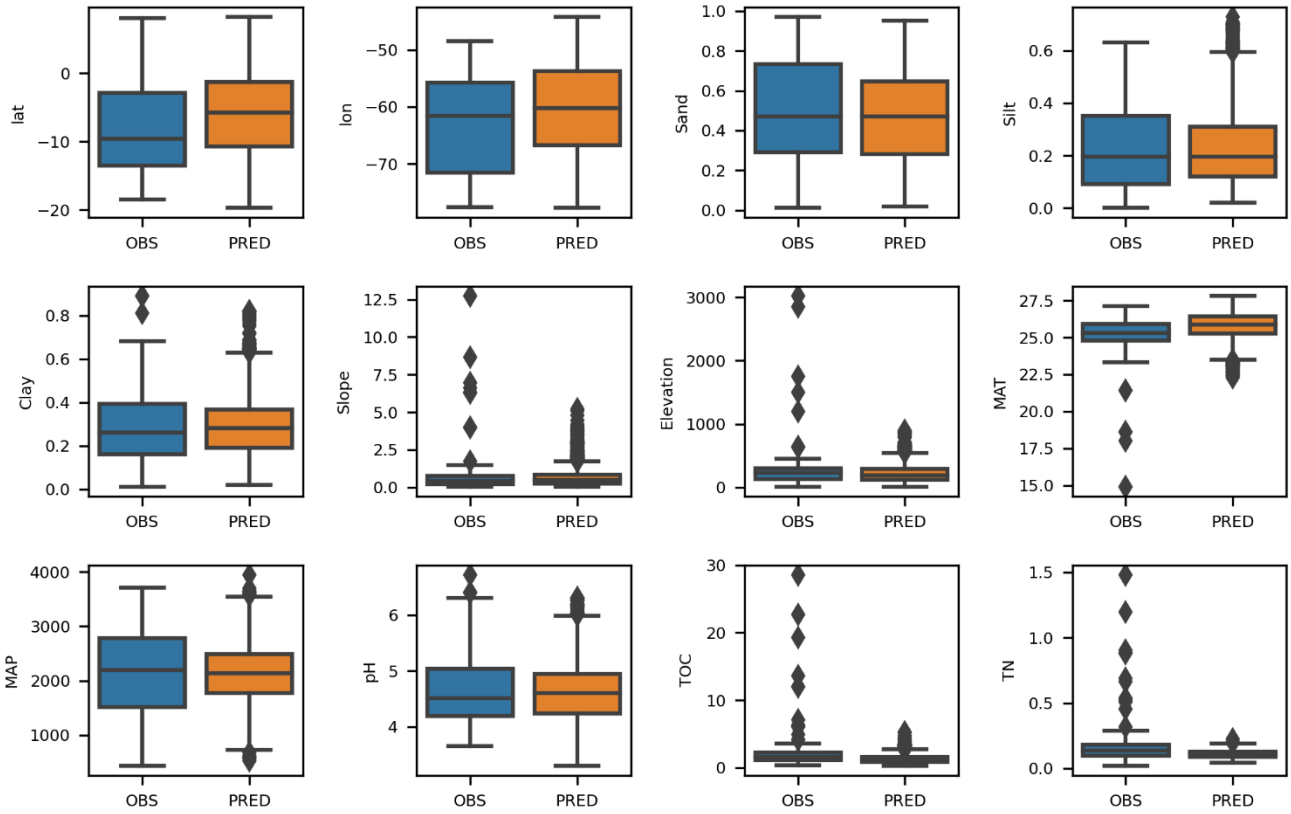
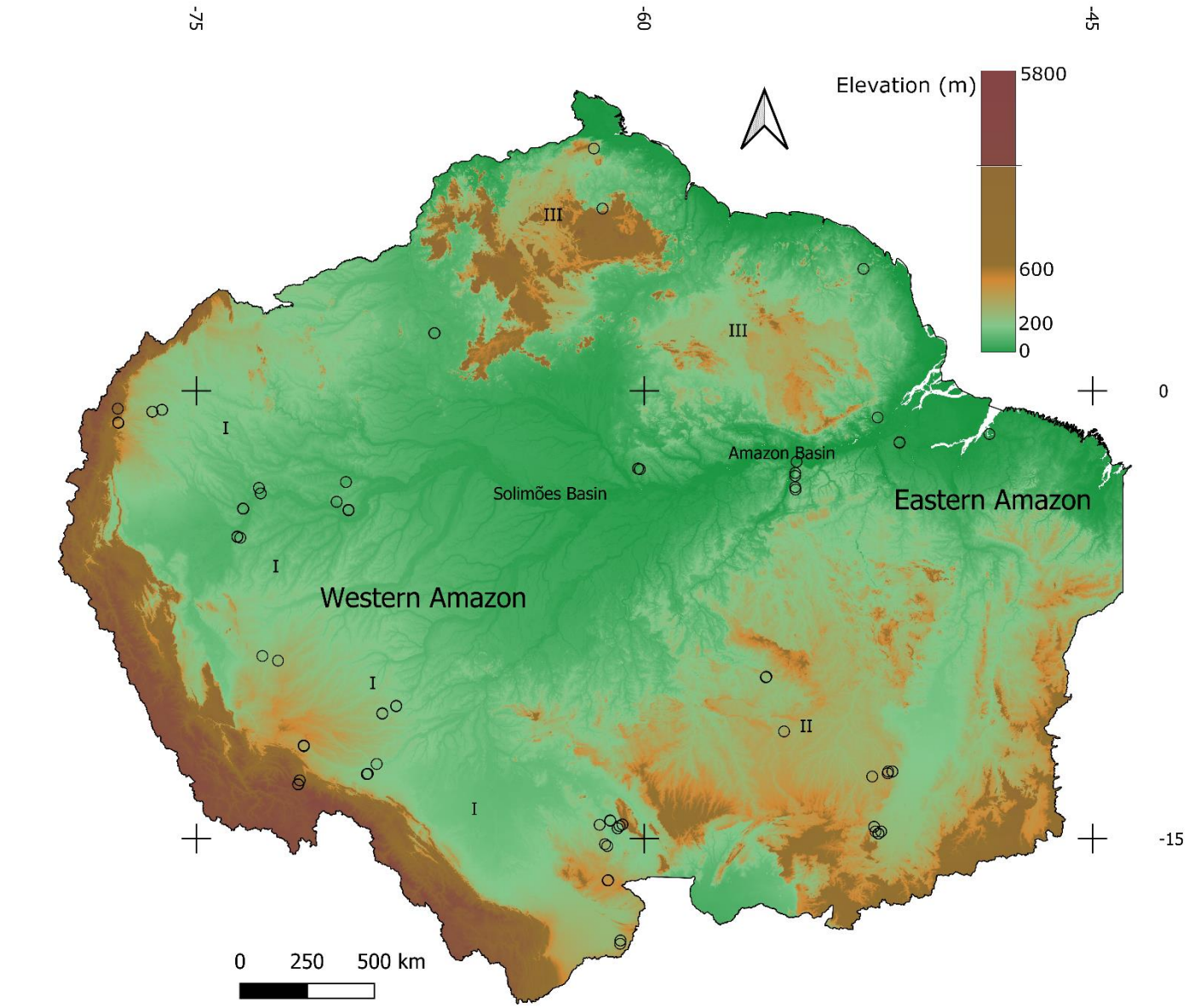
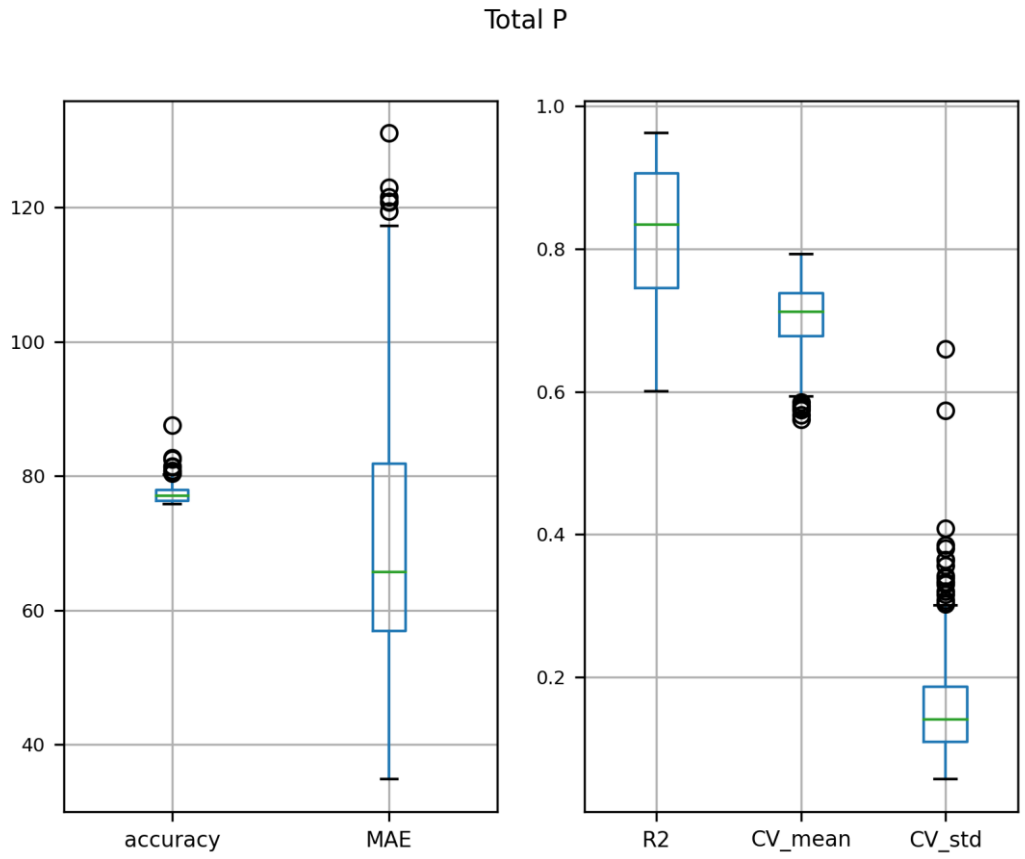


Figure S5: Distribution of the features in the *fitting dataset* (OBS) and in the *predictive dataset* (PRED) after the exclusion of the grid cells with values above the defined DI threshold.



35

Figure S6: Shuttle Radar Topography Mission digital elevation model of the study area. Source: Saatchi (2013). (I) Mark the approximate positions of the Amazonian foreland basins; (II) Brazilian shield; (III) Guiana shield. Black circles mark the locations of soil sampling locations in the fitting dataset. Solimões and Amazon sedimentary basins.



40

Figure S7: Evaluation metrics distribution for the 300 models selected to the prediction of total P concentration. Accuracy (%) (Eq. 1 in main text) and MAE – mean absolute error (mg kg^{-1}) appear in the left panel. Coefficient of determination (R^2), cross validation R^2 score (CV_mean) and cross validation standard deviation (CV_std) appear in the right panel.

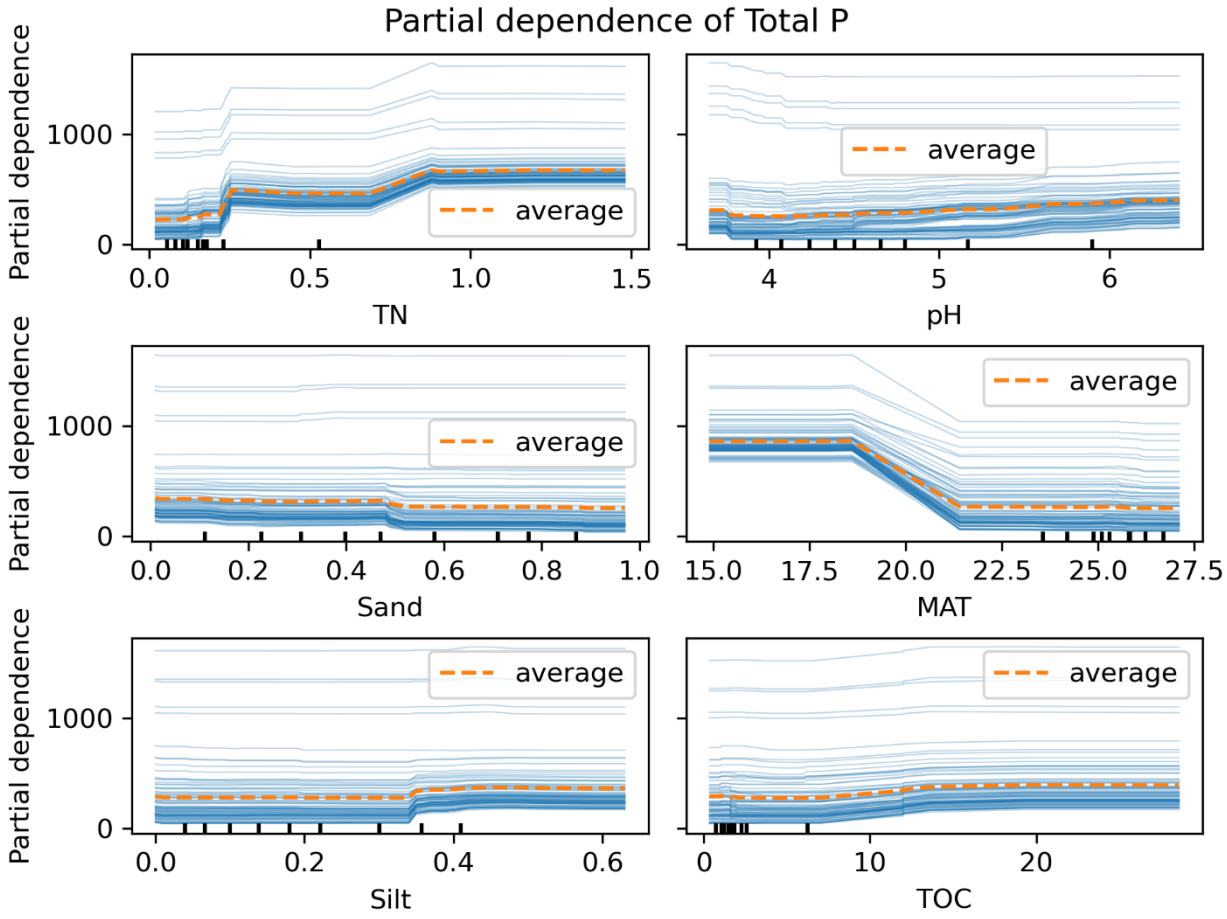


Figure S8: Individual Conditional Expectation plots showing the partial dependence of total P to the 6 most important features ranked by the mean MDA scored in the permutation importance calculation. The black markers in the features axes indicate de deciles of the training data (n=85). The partial dependence in the plot was estimated using the random forest model fitted to total P with the *best* accuracy score. Blue lines denote the partial dependence for each sample in the fitting dataset.

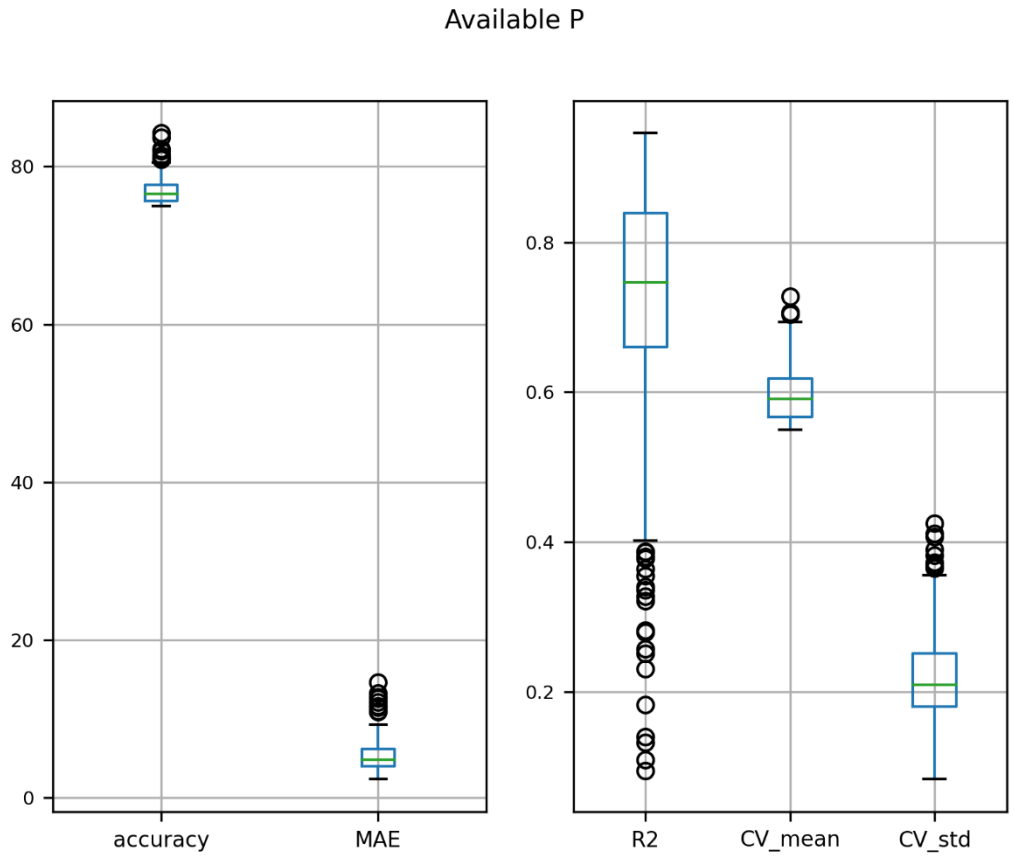


Figure S9: Evaluation metrics distribution for the 419 models selected to the prediction of available P concentration. Accuracy (%) (Eq. 1 in main text) and MAE – mean absolute error (mg kg^{-1}) appear in the left panel. Coefficient of determination (R^2), cross validation R^2 score (CV_mean) and cross validation standard deviation (CV_std) appear in the right panel.

55

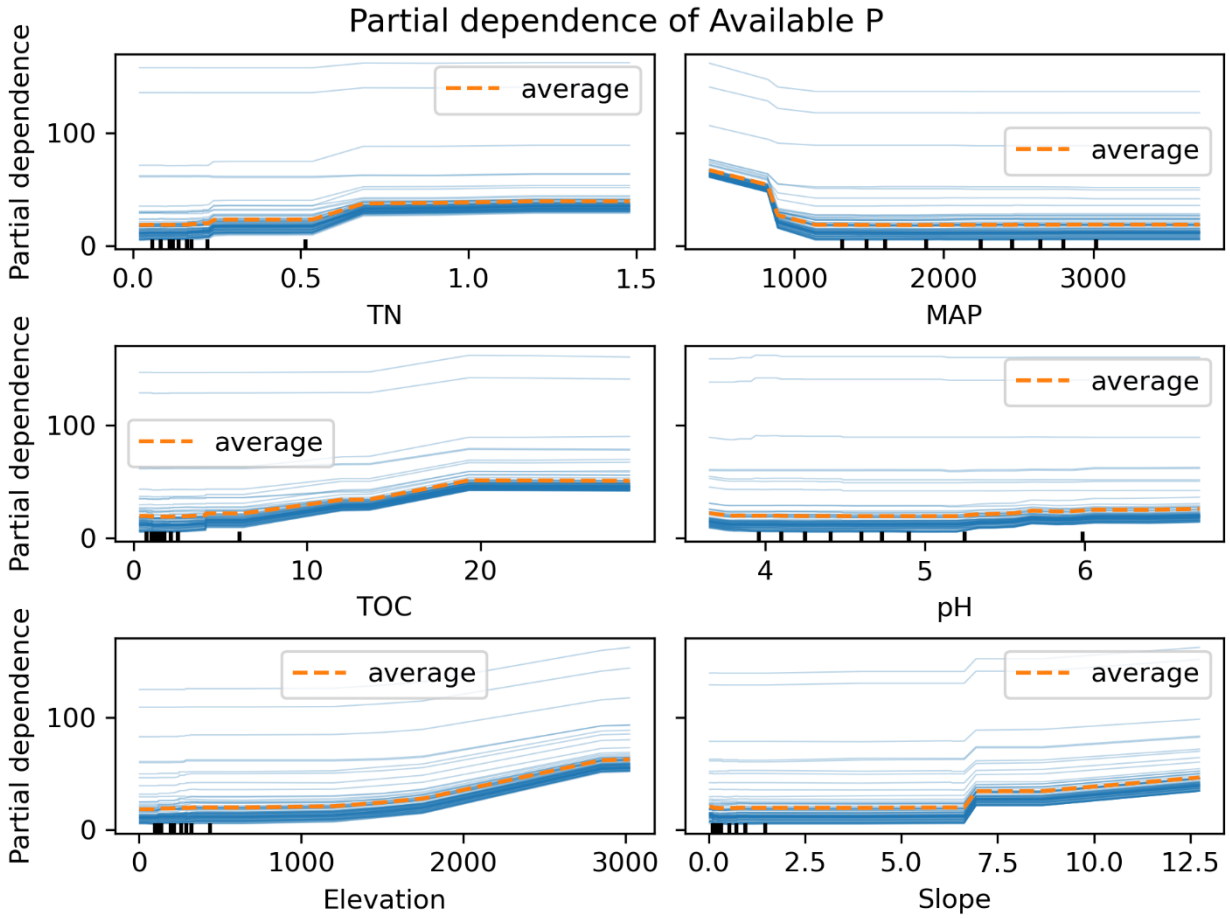


Figure S10: Individual Conditional Expectation plots showing the partial dependence of available P to the 6 most important features ranked by the mean MDA scored in the permutation importance calculation. The black markers in the features axes indicate de deciles of the training data (n=85). The partial dependence in the plot was estimated using the random forest model fitted to available P with the *best* accuracy score. Blue lines denote the partial dependence for each sample in the fitting dataset.

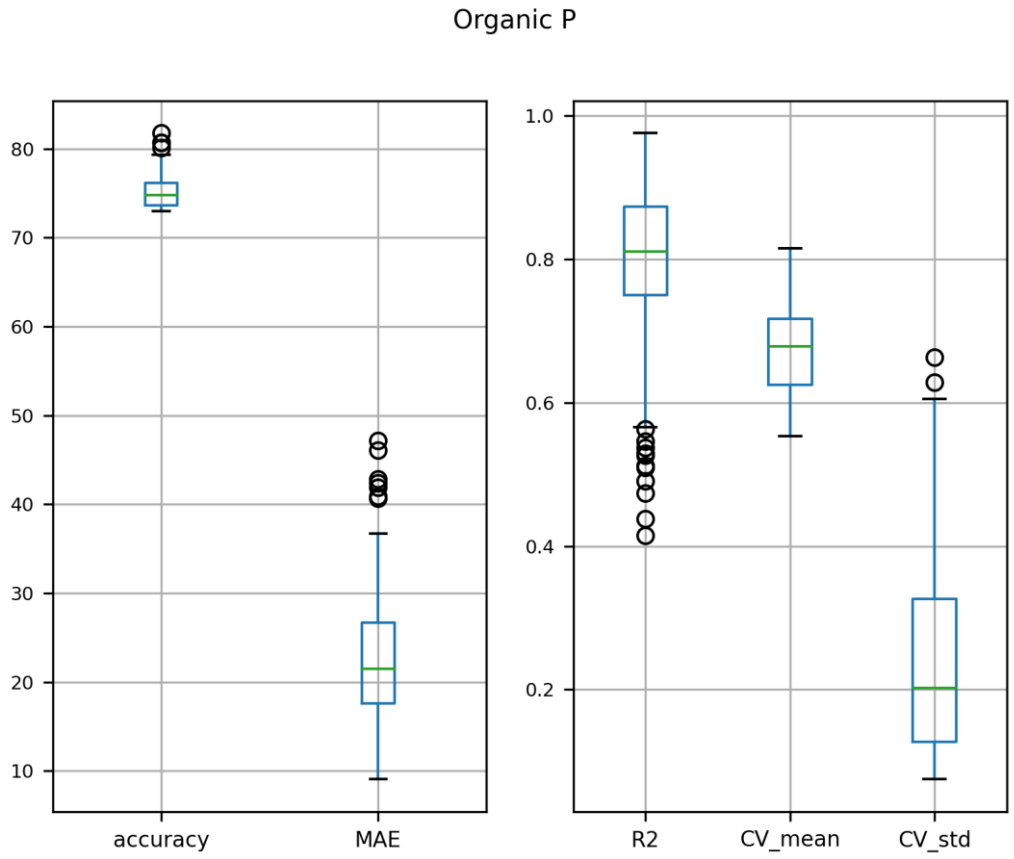


Figure S11: Evaluation metrics distribution for the 247 models selected to the prediction of organic P concentration. Accuracy (%) (Eq. 1 in main text) and MAE – mean absolute error (mg kg^{-1}) appear in the left panel. Coefficient of determination (R^2), cross validation R2 score (CV_mean) and cross validation standard deviation (CV_std) appear in the right panel.

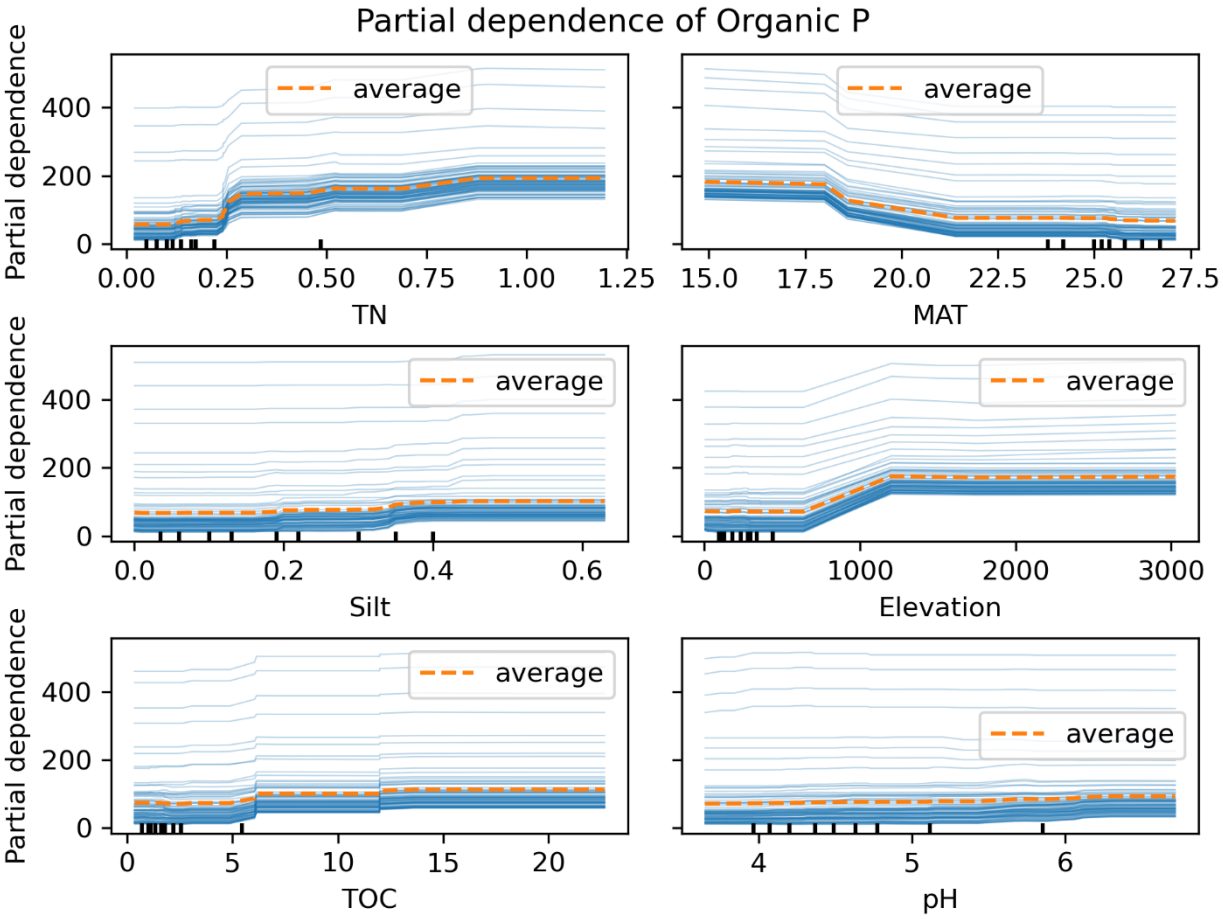


Figure S12: Individual Conditional Expectation plots showing the partial dependence of organic P to the 6 most important features ranked by the mean MDA scored in the permutation importance calculation. The black markers in the features axes indicate de deciles of the training data (n=85). The partial dependence in the plot was estimated using the random forest model fitted to available P with the *best* accuracy score. Blue lines denote the partial dependence for each sample in the fitting dataset.

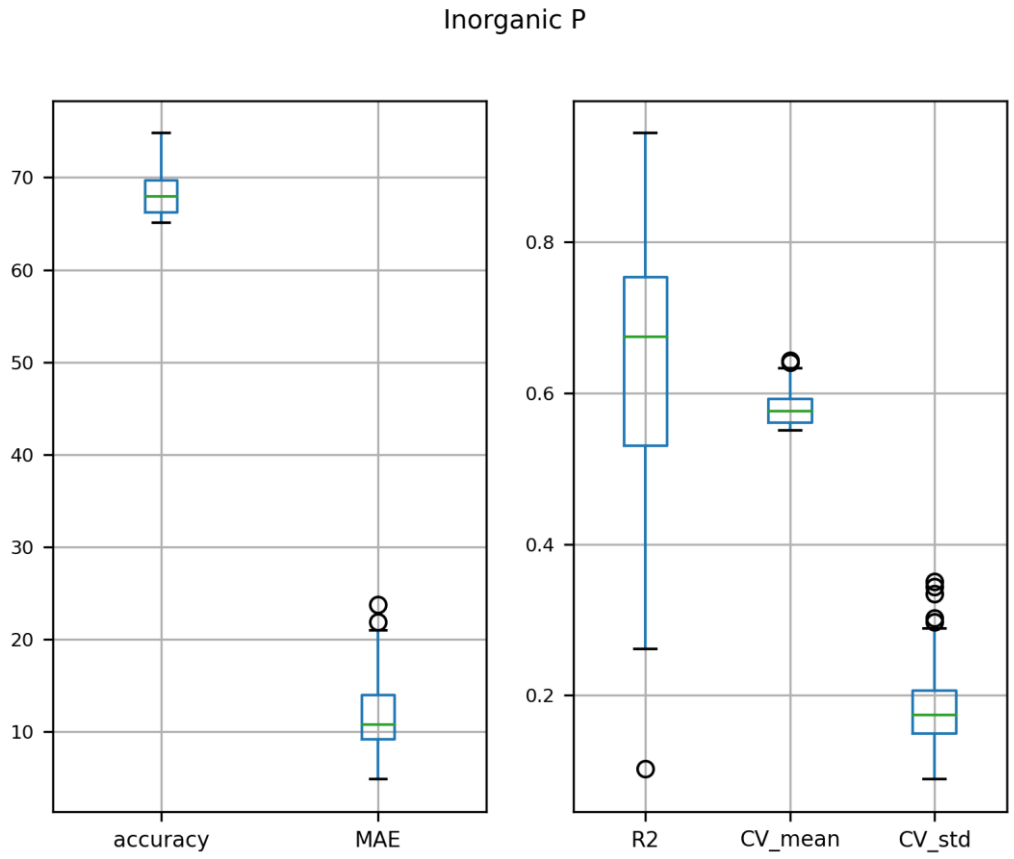


Figure S13: Evaluation metrics distribution for the 102 models selected to the prediction of inorganic P concentration. Accuracy (%) (Eq. 1 in main text) and MAE – mean absolute error (mg kg^{-1}) appear in the left panel. Coefficient of determination (R^2), cross validation R2 score (CV_mean) and cross validation standard deviation (CV_std) appear in the right panel.

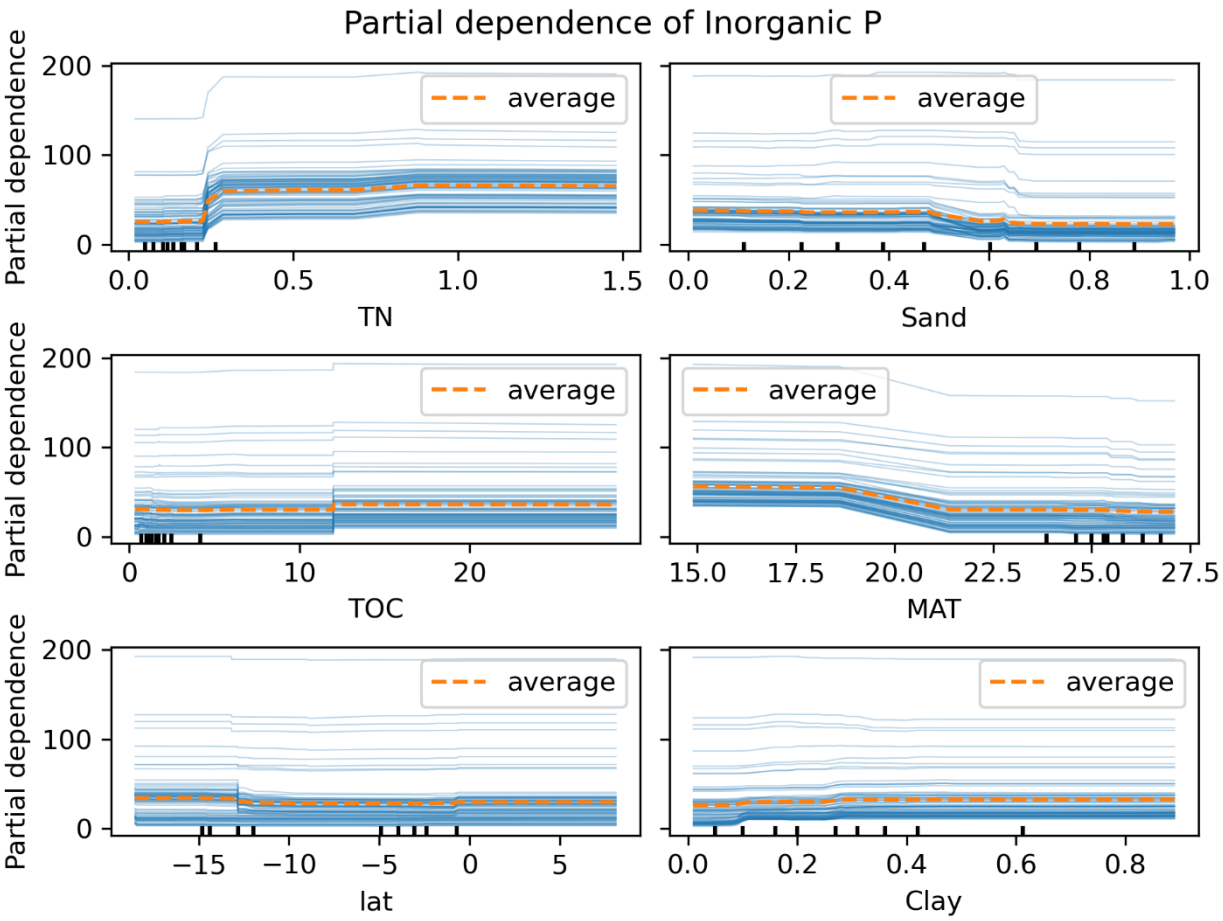


Figure S14: Individual Conditional Expectation plots showing the partial dependence of inorganic P on the 6 most important features ranked by the mean MDA scored in the permutation importance calculation. The black markers in the features axes indicate de deciles of the training data (n=85). The partial dependence in the plot was estimated using the random forest model fitted to available P with the *best* accuracy score. Blue lines denote the partial dependence for each sample in the fitting dataset.

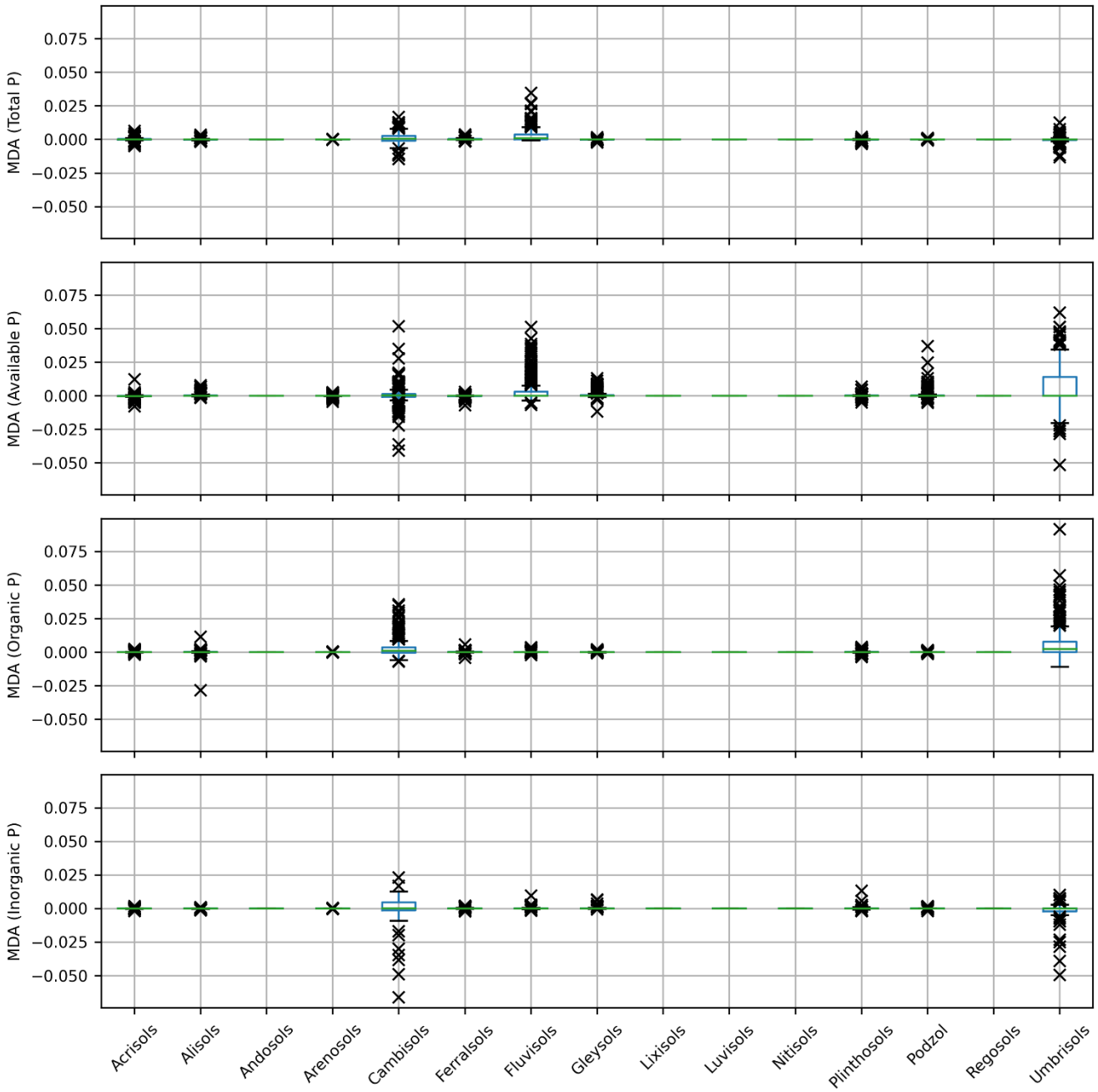
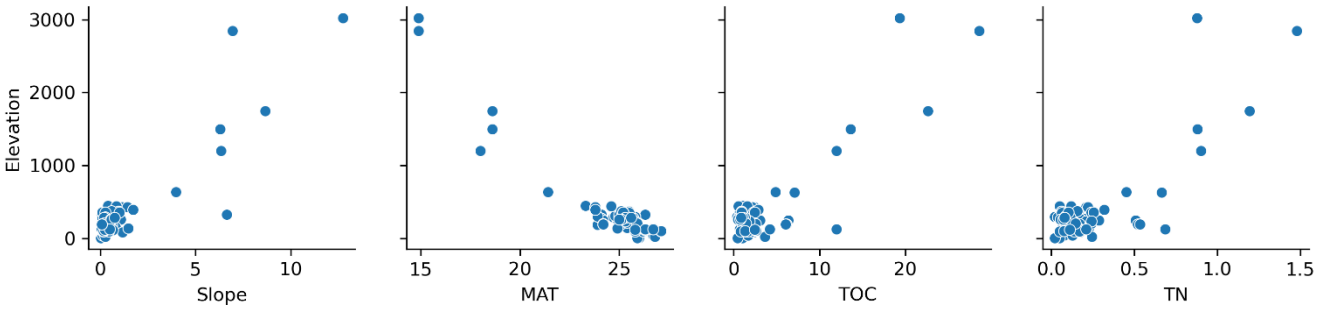


Figure S15: Soil reference groups permutation importance – or MDA (Mean Decrease in Accuracy). Distribution of means for the set of random forest models selected for each P form (Table 3). Positive (negative) values of MDA indicates that the 'exclusion' of the variable decrease (increase) the random forest model accuracy. Higher values of MDA indicate higher variable importance. Each selected model was permuted 120 times. The internal variability (Standard Deviation of MDA) of each model is not presented.



90

Figure S16: Scatterplots of Slope, MAT, TOC, and TN in relation to Elevation in the fitting dataset.

Table S1: Descriptive statistics of the target variables in the phosphorus (fitting) dataset used to train and test the random forest regression models. Units are mg kg^{-1}.

	mean	std	min	25%	50%	75%	max
Occluded	141.25	160.31	0.49	50.53	87.95	201.05	1213.27
Mineral	11.09	28.70	0.00	0.82	1.36	5.80	186.58
Inorganic P	33.81	42.51	1.00	11.68	21.96	36.34	252.60
Organic P	78.93	104.83	5.69	24.86	46.06	85.28	618.39
Available P	19.06	25.54	3.29	8.60	11.83	18.89	188.95
Total P	284.13	305.03	24.57	109.96	167.18	355.75	1966.67

95

Table S2: Descriptive statistics of the features in the phosphorus dataset used to train and test the random forest regression models. Units are presented in the main text (Table 1).

	mean	std	min	25%	50%	75%	max
lat	-8.05	6.32	-18.52	-13.49	-9.59	-2.83	8.11
lon	-64.21	8.51	-77.63	-71.54	-61.60	-55.77	-48.45
Sand	0.49	0.27	0.01	0.29	0.47	0.73	0.97
Silt	0.22	0.16	0.00	0.09	0.20	0.35	0.63
Clay	0.28	0.19	0.01	0.16	0.26	0.39	0.89
Slope	0.95	1.88	0.02	0.19	0.41	0.74	12.74
Elevation	311.96	436.51	4.00	125.30	226.85	299.47	3025.00
MAT	24.97	2.07	14.90	24.78	25.30	25.90	27.10
MAP	2153.76	741.29	433.10	1512.10	2195.00	2777.65	3710.70
pH	4.69	0.72	3.65	4.19	4.52	5.04	6.72
TOC	2.69	4.19	0.35	1.08	1.57	2.29	28.59
TN	0.20	0.23	0.02	0.10	0.14	0.18	1.48

Table S3: Descriptive statistics of the features in the predictive dataset used to predict the target P pools with the trained and tested random forest regression models. Units are presented in the main text (Table 1).

	mean	std	min	25%	50%	75%	max
lat	-5.74	6.52	-20.25	-11.13	-5.75	-0.75	9.75
lon	-61.30	8.59	-79.25	-67.75	-61.25	-54.25	-44.25
Sand	0.46	0.21	0.00	0.28	0.47	0.64	0.95
Silt	0.25	0.16	0.00	0.12	0.21	0.33	0.79
Clay	0.29	0.14	0.00	0.18	0.27	0.36	0.82
Slope	1.26	2.18	0.03	0.26	0.53	1.02	16.40
Elevation	385.58	667.73	4.24	120.71	203.91	333.64	4769.19
MAT	25.09	3.00	2.43	24.99	25.77	26.41	27.81
MAP	2124.96	603.52	398.08	1722.44	2118.63	2494.02	4657.22
pH	4.68	0.68	3.20	4.24	4.61	5.03	8.21
TOC	1.48	1.71	0.27	0.88	1.18	1.67	35.18
TN	0.11	0.03	0.24	0.09	0.11	0.13	0.54

Table S4: Occurrence of soil reference groups in the predictive dataset (Pred.) as a count of grid cells for each group. Excluded cells after the calculation of the dissimilarity index (Excl. DI). Number of observations in the fitting dataset (Fit.) presents the counts of in situ measurements in the phosphorus dataset for each soil class. Undefined grid cells (Undef.) in the predictive dataset since Histosols, Leptosols, Solonetz, and Solonchak do not appear in the fitting dataset, but are in the predictive dataset. The classes Andosols and Umbrisols are absent in the predictive dataset because of the spatial aggregation process.

RSG	**Pred.**	**Excl. DI**	**Fit.**	**Undef.**
Acrisols	727	43	13	-
Alisols	5	-	10	-
Andosols	-	-	1	-
Arenosols	99	14	8	-
Cambisols	187	68	18	-
Ferralsols	909	54	26	-
Fluvisols	15	-	3	-
Gleysols	189	44	5	-
Histosols	4	4	-	4
Leptosols	172	115	-	172
Lixisols	42	5	1	-
Luvisols	5	4	1	-
Nitisols	1	-	1	-
Plinthosols	256	17	11	-
Podzols	44	15	5	-
Regosols	92	68	1	-
Solonchak	1	-	-	1
Solonetz	1	-	-	1
Umbrisols	-	-	4	-

References

Meyer, H. and Pebesma, E.: Predicting into unknown space? Estimating the area of applicability of spatial prediction models, Methods Ecol Evol, 12, 1620-1633, 10.1111/2041-210x.13650, 2021.

Saatchi, S. S.: LBA-ECO LC-15 SRTM30 Digital Elevation Model Data, Amazon Basin: 2000, ORNL Distributed Active Archive Center [dataset], 10.3334/ORNLDAAC/1181, 2013