Electronic Supplementary Information (ESI)

Predicted Losses of Sulfur and Selenium in European Soils Using Machine Learning: A Call for Prudent Model Interrogation and Selection

Gerrad D. Jones,[a*] Logan Insinga,[a] Boris Droz,[b,c] Aryeh Feinberg,[d] Andrea Stenke,[e,g] Jo Smith,[f]

Pete Smith,[f] Lenny H. E. Winkel [e,g*]

a. Department of Biological & Ecological Engineering, Oregon State University, Corvallis, Oregon, 97331, United States.

b. School of Biological, Earth and Environmental Sciences, University College Cork, Cork, Ireland.

c. Water and Environment Research Group, Environmental Research Institute, Lee Road, University College Cork, Cork, Ireland.

d. Institute for Data, Systems, and Society, Massachusetts Institute of Technology, Cambridge, MA 02139, United States.

e. Institute of Biogeochemistry and Pollutant Dynamics, ETH Zürich, 8092 Zürich, Switzerland.

f. Institute of Biological and Environmental Sciences, School of Biological Sciences, University of Aberdeen, Aberdeen AB24 3UU, United Kingdom.

g. Eawag, Swiss Federal Institute of Aquatic Science and Technology, 8600 Dübendorf, Switzerland.

* Corresponding author. e-mail: gerrad.jones@oregonstate.edu, lwinkel@ethz.ch

This ESI includes 25 pages with texts, 2 tables and 13 Figures.

Predictive variables and data processing

Table S1. Predictive variables (n = 46) governing S and Se concentrations in soil collected for the purpose of the study.

Variable	Unit	Scenario	Year coverage	Res. (km)	Ref.
Vegetation variables					
Canopy height	m		2004	0.5	[1]
Maximum green vegetation fraction	unitless		2012	1	[2]
Soil physical/chemical properties					
Sand, silt, & clay[a,b]	%		1990-2016	0.25	[3]
Bulk density	kg/m^3		1990-2016	0.25	[3]
Soil pH[a,b]	unitless		1990-2016	0.25	[3]
Cation exchange capacity (CEC)	cmol/kg		1990-2016	0.25	[3]
Soil chemical properties					
Soil organic carbon[a,c]	t C/ ha	current	1990–2000	1	[4]
		future (A1FI)	2080		
Climate					
Average daily precipitation (Precip)[a,c]	mm/yr	current	1971–2000	12.5	[5]
		future (RCP 8.5)	2071–2100		
Average daily evapotranspiration (ET)[a,c]	mm/yr	current	1971–2000	12.5	[5]
		future (RCP 8.5)	2071–2100		
Current and future aridity index (AI = PET/Precip)[a,c]	unitless	current	1971–2000	12.5	[5]
		future (RCP 8.5)	2071–2100		
Current and future evaporative index (EI = ET/Precip)[a,c]	unitless	current	1971–2000	12.5	[5]
		future (RCP 8.5)	2071–2100		
Other					
Bedrock depth	m		1990-2016	0.25	[3]
Lithology (14 var.)[d]	class			111	[6]
Population density			2000	1	[7]
Elevation[d]	m			2	[8]
Aspect, slope[d]	degree			2	derived from elevation

Mineralogy					
Chemical index of alteration (CIA)[a,b,c]	%			point	[9]
Soil metal oxide content (10 var.; SiO_2,TiO_2,Al_2O_3,…)[c]	wt % or mg/kg			point	[9]
Emission/deposition					
Total S and Se deposition[a,d]	mg S /(m^2 yr)	current	2005–2009	~310	[10]
		future (SSP5–8.5)	2095–2099		

[a] indicates variables retained for modeling, [b] indicates variables that were held constant in future predictions, [c] indicates variables collected/modeled specifically from Europe and [d] indicates when original data were in degree and converted to the equivalent in km at the equator. Res. and Ref. stand for resolution and references, respectively.

The GEMAS dataset was originally divided into grazing and plowed soils, these two designations were test as predicting variables. However, these designations were defined as not important for describing S and Se element concentrations during the selection of the most relevant predictive variable. Therefore, the soil data were pooled in the final analyses.

All element data and predictive variables within a 111 km pixel centered at the nearest 55.5 km were averaged in ArcMap 10.6 using the "Resample" tool to reduce the influence of errors and/or outliers within the datasets. Consequently, pixels containing <3 data points were removed from the analysis. All data were tested for normality by assessing skew. If |skew| <1, then the variables were considered normally distributed.[11] Without transformation, skew ranged from 0 to 14. Sulfur concentration was the most positively skewed variable, and following log10 transformation, the skew for S was 1.6, which was an improvement and deemed suitable for analysis. Although many ML techniques do not need normal distributed data, normalizing the data reduces the influence of outliers and can increase prediction efficiency and function convergences.[12] Finally, because support vector machines are sensitive to variations in scale,[13] all predictive variables were z-score transformed (i.e., $z = (x-\bar{x})/stdev$, where x is the sample observation, $\bar{x}$ is the sample mean, and stdev is the sample standard deviation). This normalized the data such that the mean and standard deviation were 0 and 1, respectively. For variables with future data, the future data were also scaled based on the mean and standard deviation of the current conditions.

We strove to select only the most relevant predictive variables using various methods to improve interpretability and reduce overfitting. We used linear support vector regression (Lin-SVR) analysis to pre-screen variables with low explanatory power. SVR models were set up as described below in "Machine learning analysis". The coefficient weight (estimator.coef_) of all variables was retained, and those with coefficients near 0 were removed. Additionally, model sensitivity analyses were used to evaluate the independent effect of each variable in the Lin-SVR analysis (see "Machine learning analysis" below). Predictive variables with disproportionately low importance or eliciting a small response in soil element concentrations were removed from analyses. Variables with low coefficient weights also exhibited little change in the sensitivity analyses.

Mechanistically, all retained variables are known to govern the retention of S and Se. These mechanisms are largely sources, sink, and transport mechanisms. Total deposition are sources of both elements to soils.[10] As AI increases, soils become drier and thus oxidizing.[14, 15] Oxidized species of S and Se are more mobile,[16, 17] so we expect a negative relationship between AI and both elements. Increases in clay content are known to increase S and Se sorption.[18, 19] Increases in pH are known to decrease S and Se sorption.[18, 20] ET could influence soil element concentrations through multiple mechanisms. For example, ET reduces water from the soil column, which thus reduces the potential for transport.[21] Furthermore, ET can contribute to the removal of trace elements from the soil through plant uptake.[22] It is even possible that plants can pump trace elements from the subsurface to the surface through root uptake,[23, 24] incorporation into leaf tissues,[25] and deposition and decomposition of leaves.[25] Finally, precipitation is likely the most complex variable and can influence trace element concentrations in soils through multiple mechanisms. For example, precipitation influences evapotranspiration,[26] soil organic matter development,[27] soil pH,[28] pedogenesis,[29] as well as

deposition to the soil surface and transport (i.e., leaching) away from the surface.[30] While Jones et al.[31] found a total positive effect driven by precipitation, the independent effect was negative, which is potentially an indication that precipitation is a driver of leaching. In general, we expect positive relationships between S and Se with deposition, clay, CIA, and SOC, and negative relationships with AI and pH. ET and precipitation are difficult to predict a priori, but based on Jones et al.,[31] we expect a positive relationship with ET and a negative relationship with precipitation.

Machine-learning analysis

Our objective was not to compare the performance of the different machine-learning (ML) techniques to identify the technique with the highest performance but to develop models that generate soil S and Se predictions that are consistent with literature regarding known mechanisms. Because we did not know which technique s would perform well, we chose four technique s for their diversity. Each technique has a variety of settings and tuning parameters. The values and ranges of tuning parameters for each model are documented in the code and Fig. S1 to S4 below. For any adjustable parameter not listed, default values were used. Figures illustrate how sensitive model performance (RMSE) was to changes in each parameter. In some instances, the model performance was irresponsive to changes in a particular variable. Cross validation was performed using a 80%:20% split for training and testing datasets. Each split was selected randomly for each iteration (n = 100 for each model). The tuning parameters that resulted in the lowest RMSE were chosen as the best values. The top 10 values were colored cyan (best value) and red (top 2-10 values). The same ranges of values used for both S and Se models, but only results for Se are illustrated below.

Neural network/multilayer perceptron parameters

neural_network.MLPRegressor((hidden_layer_sizes=hidden_layer_sizes, activation='tanh', solver='adam', alpha=alpha, batch_size='auto', learning_rate='constant', learning_rate_init=0.001, power_t= power_t, max_iter=5000, shuffle=True, random_state=None, tol=0.001, verbose=False, warm_start=False, momentum=0.9, nesterovs_momentum=True, early_stopping=False, validation_fraction=0.1, beta_1=beta_1, beta_2=beta_2, epsilon=epsilon, n_iter_no_change=2)

Parameter ranges for all tuning variables

hidden_layer_sizes = np.linspace(1,51,50,endpoint=True, dtype=int)
alpha = np.geomspace(0.00000001, 100,num=1000, endpoint=True)
power_t = np.geomspace(0.00000001, 100,num=1000, endpoint=True)
beta_1 = np.geomspace(0.00000001, 0.99999999, num=1000, endpoint=True)
beta_2 = np.geomspace(0.00000001, 0.99999999, num=1000, endpoint=True)
epsilon = np.geomspace(0.0000000001, 100, num=10000, endpoint=True)

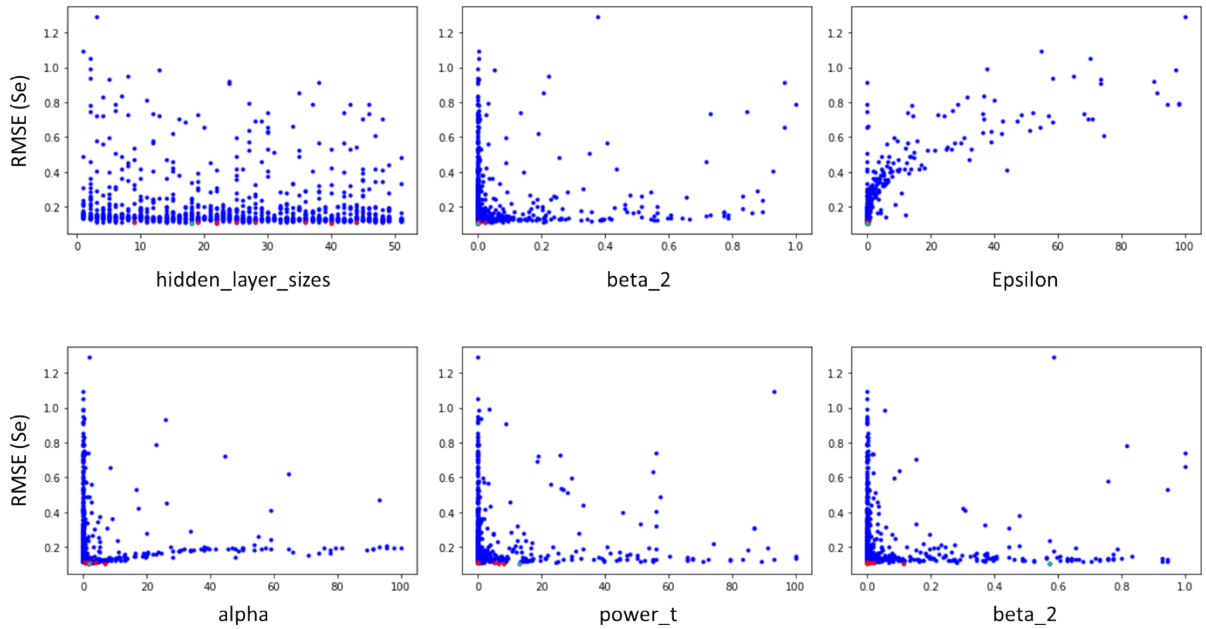

Fig. S1. With the Se data illustrated above, the MLP model performance was most sensitive to changes in epsilon.

Support vector regression (linear kernel) parameters

svm.SVR(kernel='linear', C=C, epsilon=epsilon, gamma=1, tol=0.001, shrinking=True,
cache_size=200, verbose=False, max_iter=-1)

Parameter ranges for all tuning variables
C = np.geomspace(0.0001, 10,num=1000,endpoint=True)
epsilon = np.geomspace(0.00001, 200, num=1000, endpoint=True)

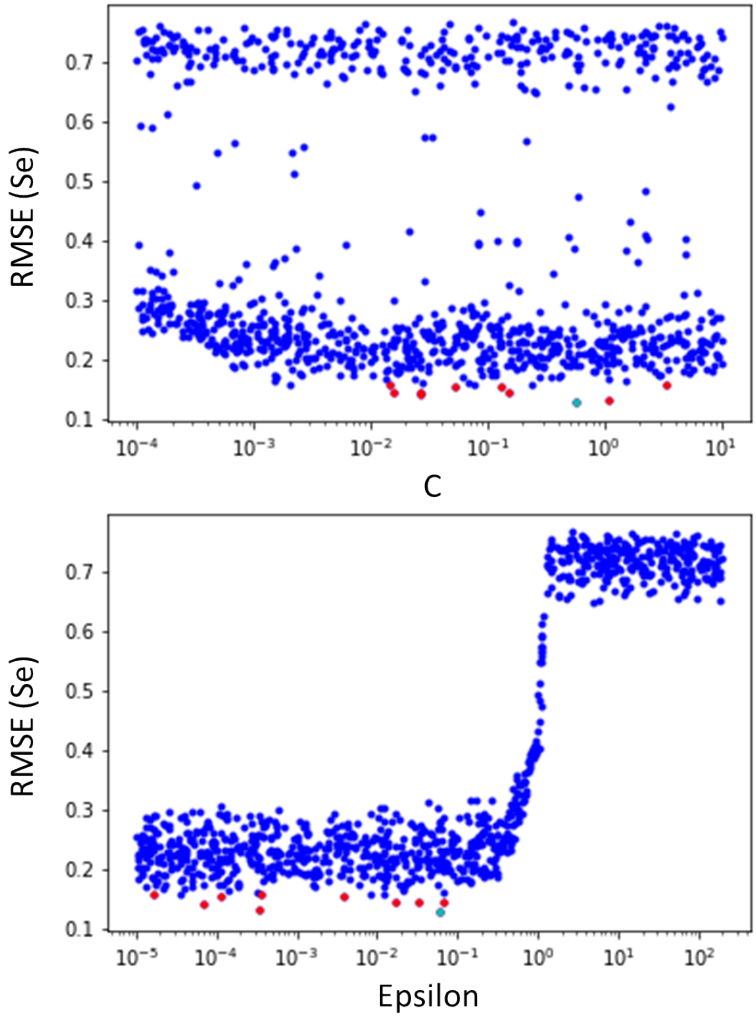

Fig. S2. With the Se data, the Lin-SVR model was performance was most sensitive to changes in epsilon. C values below 10^{-2} resulted in the best performance.

Support vector regression (radial basis function kernel) parameters

svm.SVR(kernel='rbf', gamma=gamma, tol=0.001, C=C, epsilon=epsilon, shrinking=True, cache_size=200, verbose=False, max_iter=-1)

Parameter ranges for all tuning variables

C = np.geomspace(0.0001, 10,num=1000,endpoint=True)

gamma = np.geomspace(1,1,num=1000,endpoint=True)

epsilon = np.geomspace(0.00001,200, num=1000, endpoint=True)

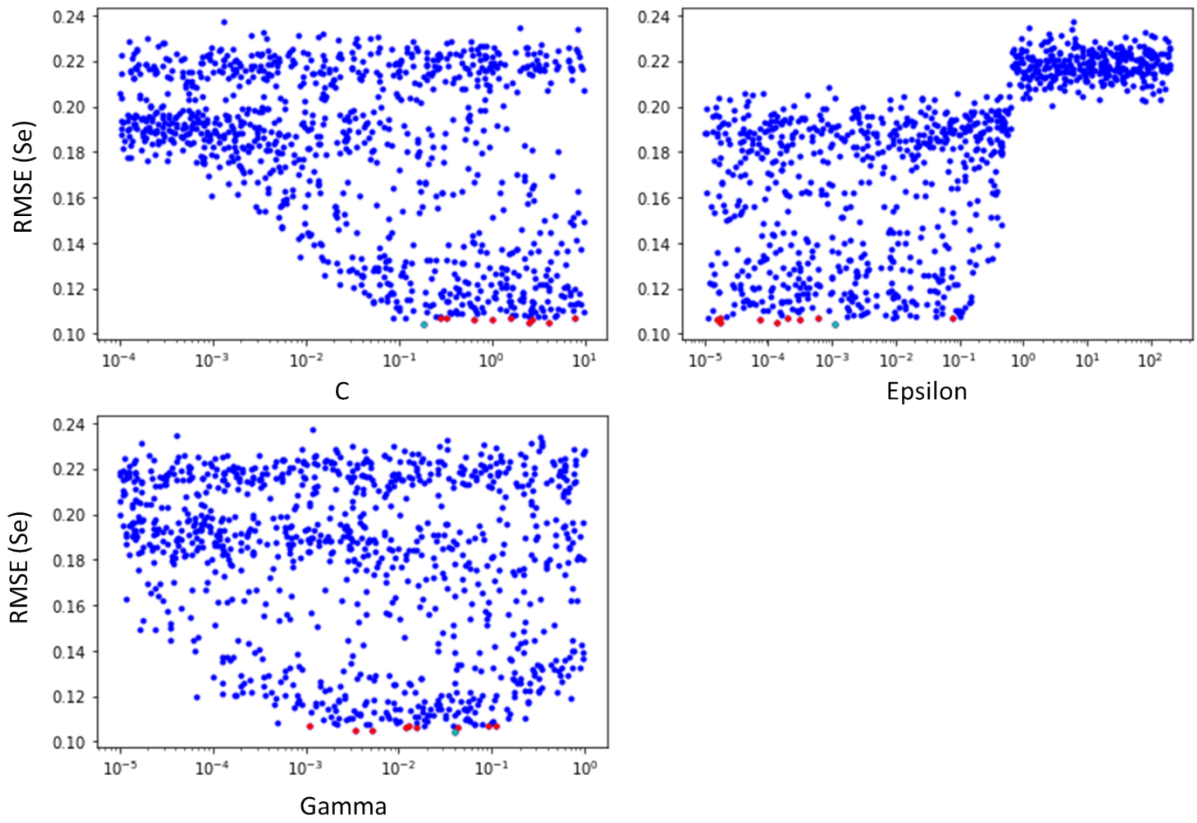

Fig. S3. With the Se data, the RBF-SVR model was performance was equally sensitive to changes in epsilon, C, and gamma.

Random forest regression parameters

ensemble.RandomForestRegressor((n_estimators=n_estimators, criterion='mse',
max_depth=max_depth, min_samples_split=min_samples_split,
min_samples_leaf=min_samples_leaf, min_weight_fraction_leaf=0.0,
max_features='sqrt',max_leaf_nodes=None, min_impurity_decrease=0.0,
bootstrap=True, oob_score=False, n_jobs=1, random_state=None,
verbose=0,warm_start=False)

Parameter ranges for all tuning variables
n_estimators = np.arange (10,1000,1)
max_depth = np.arange (2,100,2)
min_samples_split_ = np.arange (2,50,1)
min_samples_leaf_ = np.arange (2,50,1)

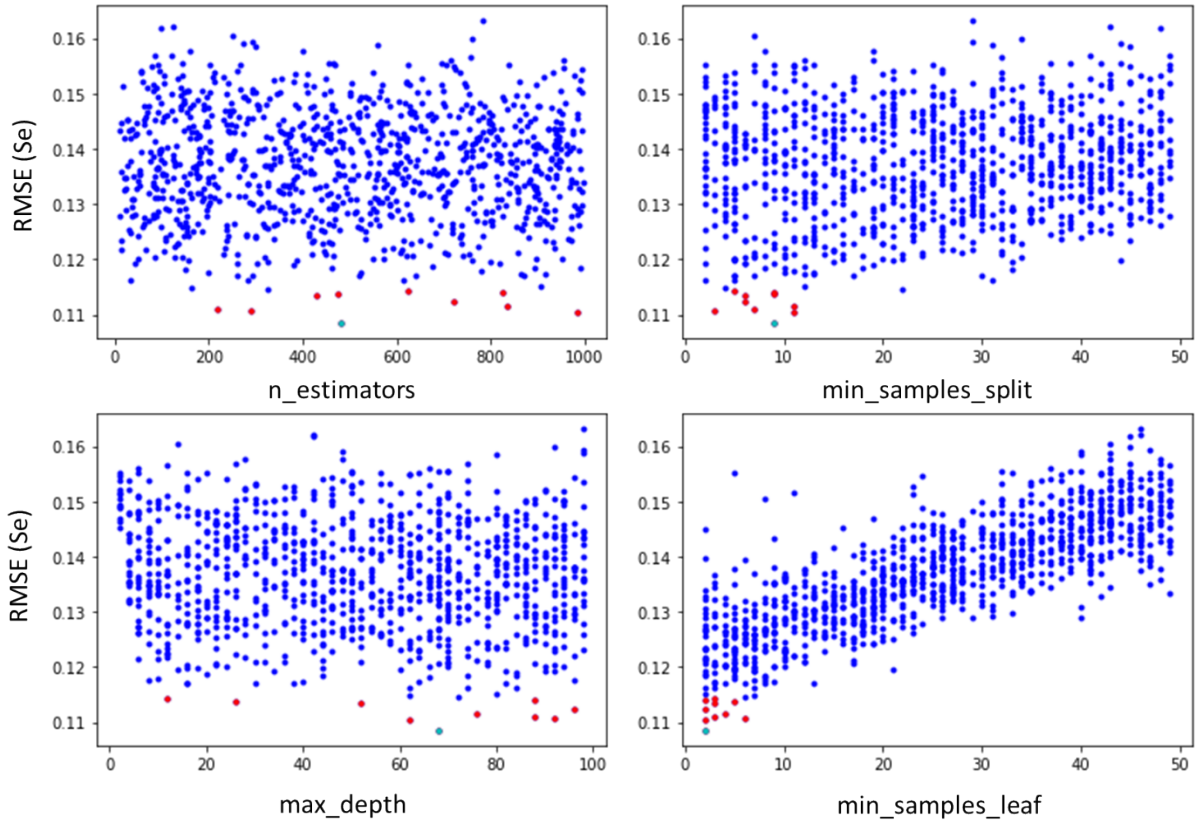

Fig. S4. With the Se data, the random forest model performance was most sensitive to min_samples_leaf. To a lesser extent, max depth and min_samples_split were also drivers of importance.

Model overfitting, accuracy, and precision calculations

Once the optimum tuning parameters were identified, the models were iterated 100 times to evaluate overfitting and to generate an average (i.e., ensemble) and standard deviation of the final models. Like model tuning, the training and testing datasets were chosen at random during each iteration. The average training and testing performance (R^2) was recorded, and for each iteration, our prediction was made for all pixels. While there is no threshold to indicate the presence of overfitting, an optimal model presents the lowest difference, close to zero, between the training and testing performance (i.e., $\Delta R^2 = R^2_{train} - R^2_{test}$). Overfitting could be considered when ΔR^2 is high, here an arbitrary threshold of ≥0.2is considered. Inversely underfitting is observed when the ΔR^2 is negative (< -0.2). Accuracy is a measure of how close the modeled estimate is from the observed value, and precision is a measure of the variability around a modeled estimate. Some pixels were over or under predicted, which is potentially the result of specific sources or mechanisms that operate locally (e.g., mining activity). A limitation of any modeling is not having all variables to fully describe all processes controlling element concentrations, and for continental-scale modeling, some data are simply not available (e.g., specific anthropogenic sources). Thus, we felt it necessary to interpret only those pixels that were well characterized by the model. Although we excluded pixels from interpretation, these data points were not excluded from the model development to avoid influence data bias, which could artificially skew the mathematical relationships during model development. Pixels that failed accuracy and precisions filters were not interpreted for future analyses, but the data were retained within the model to make future predictions.

Sensitivity analyses and partial dependence plots

After each model iteration during the final run, the models predicted element concentrations for a hypothetical set of predictive variable values whereby one predictive variable was allowed to vary across the entire domain of the dataset while all others were held constant at their mean value.[31] Because each predictive variable was transformed to z-scores, the mean value of each variable was 0. Across all predictive variables, the spread of the dataset ranged from approximately −4.4 to +6.1 standard deviations of its mean. This range was divided into 26 intervals each with a difference of 0.4 standard deviations (i.e., −4.8, −4.4, −4.0,..., 6). The output of the univariate sensitivity analysis was trimmed such that predictions did not extend beyond the domain of any predictive variable. By varying a single parameter and holding all others constant, we were able to evaluate the independent effects of each variable by accounting for the effects of the others. Partial dependence plots were created using sklearn.inspection.PartialDependenceDisplay.from_estimator. All default settings were used. As expected, the results of the sensitivity and partial dependence analyses were similar.

Model residuals

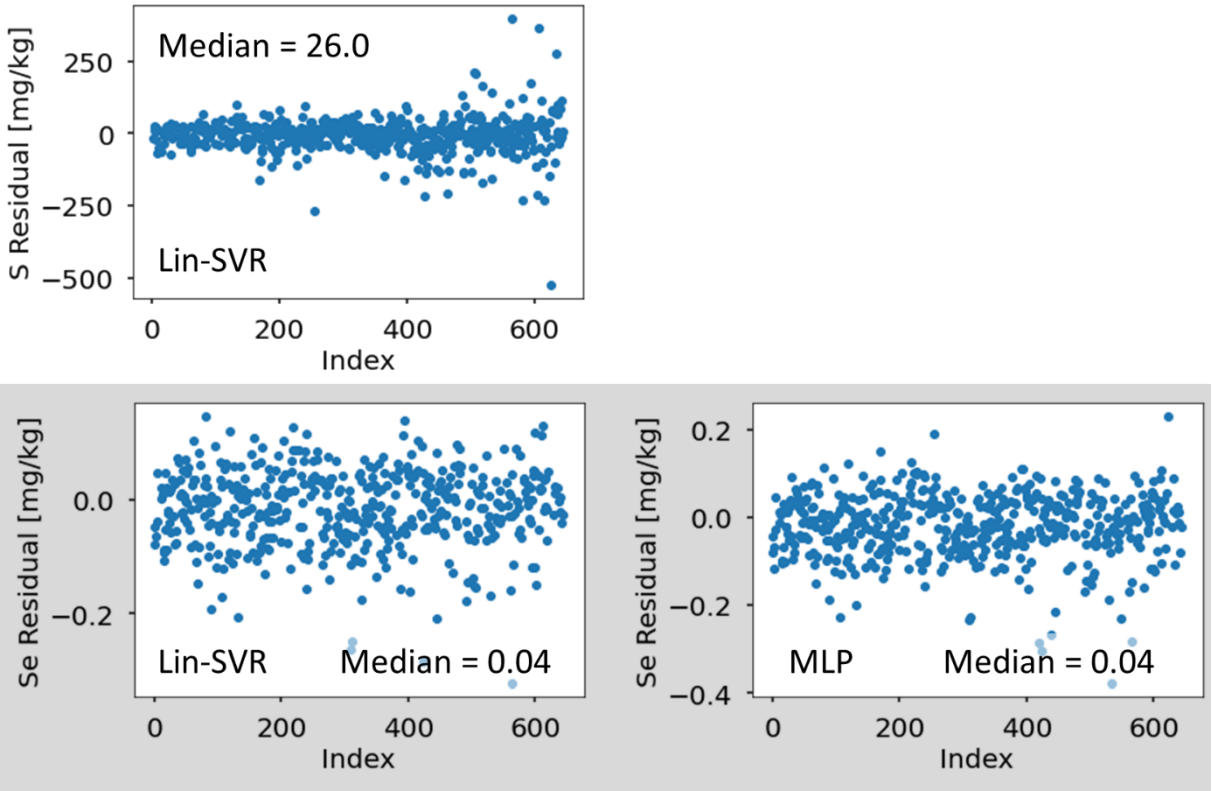

Fig. S5. Residual of the model for sulfur (S) and selenium (Se) prediction. Abbreviations include linear support vector regression (Lin-SVR) and multilayer perceptron (MLP).

Poor model performance

Table S2. Average performance (R2) of machine-learning (ML) models. Regression models include support vector regression (SVR) with a radial basis function (RBF) and linear kernels, multilayer perceptron (MLP), and random forest. Cross validation was performed using a 80%:20% split for training and testing datasets. Each split was selected randomly for each iteration (n = 100 for each model).

	S			Se		
Model	Train	Test	ΔR^2	Train	Test	ΔR^2
RBF-SVR	0.66	0.57	0.09	0.70	0.66	0.04
Lin-SVR	0.52	0.55	–0.03	0.64	0.62	0.02
MLP	0.61	0.57	0.04	0.66	0.64	0.02
RFR	0.80	0.60	0.20	0.84	0.63	0.21

$\Delta R^2 = R^2_{train} - R^2_{test}$

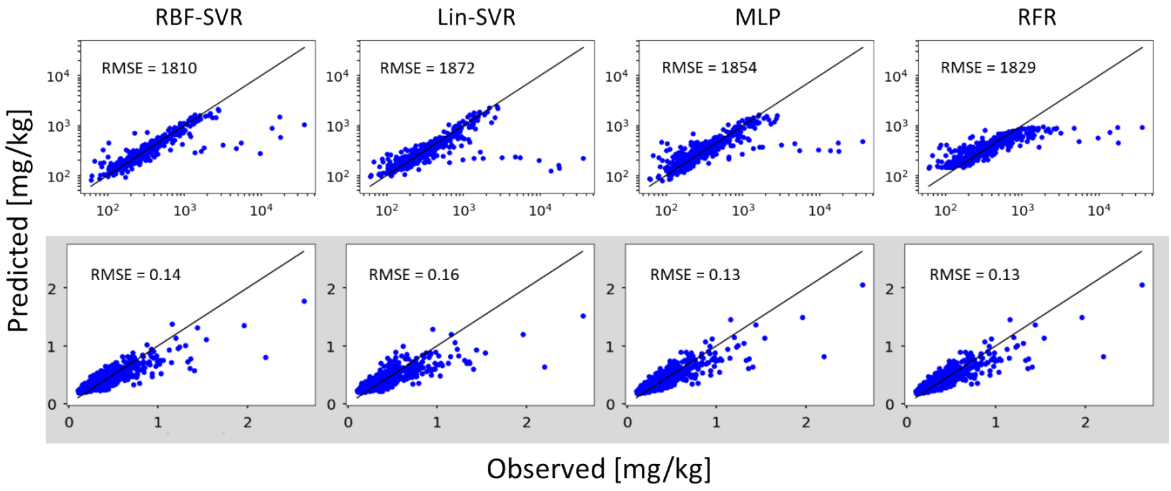

Fig. S6. Observed vs predicted concentrations of all four machine-learning (ML) models for S and Se (gray box). All data are based on average values (n = 100). Abbreviations include sulfur (S), selenium (Se), support vector regression (SVR), multilayer perceptron (MLP), root mean squared error (RMSE). The RMSE was calculated based on the final average prediction (illustrated in Fig. S6) instead of the average of each prediction (as presented in Table S1).

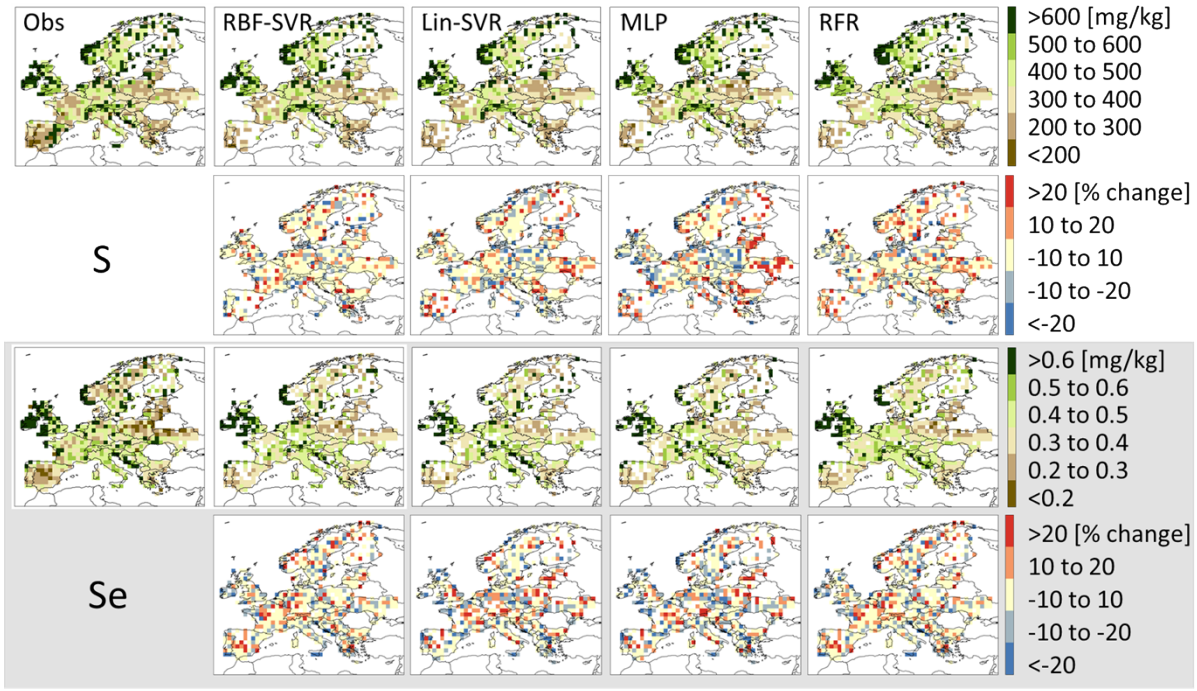

Fig. S7. Geographic distribution of observed (Obs) and modeled soil S and Se (gray box) concentrations and relative residuals (i.e., % change). Panels in each column correspond to their heading. Abbreviations include sulfur (S), selenium (Se), radial basis function kernel for support vector regression (RBF-SVR), linear kernel for support vector regression (Lin-SVR), multilayer perceptron (MLP), and random forest regression (RFR). Pixels that exceed accuracy and precision thresholds were excluded. The pixel resolution of each panel is 111 km.

Mechanistic and Model Evaluation

In the univariate sensitivity analyses, the element response to changes in each predictive variable was helpful for determining whether the model is capturing expected element fate and transport mechanisms in the model. While some predictive variables have multiple mechanisms that can affect element fate and transport (e.g., precipitation),[31] modal relationships between predictive variables and elements are unlikely on broad scales because it implies that the importance of one mechanism overtakes that of a second in the middle of the variable's gradient. Therefore, modal patterns may be a sign of overfitting within a model. S and Se were most sensitive to changes in SOC (Fig. S7, S10 & S11), which is not surprising given that SOC strongly increases soil retention of both elements.[18, 20] In the RBF model, the relationship between S and SOC is unimodal. For aridity index (AI), we expect S and Se concentrations to be negatively related to AI because long term warming will decrease the redox capacity of soil organic matter,[32] which might reduce soil S and Se sequestration. Therefore, we expect monotonic decreases in S and Se concentrations with high AI, however, unimodal patterns were observed for RBF-SVR for both elements and the MLP model for S, indicating that these models are inconsistent with our mechanistic expectations. Similarly, we expect increased S and Se sorption with increased CIA, resulting from increased binding sites of weathered soils, and increased clay. This trend was mostly observed for clay, but was not observed for CIA in the RBF-SVR, MLP, and RFR models for S. Both elements were largely invariant to changes in pH for all models. We expected increased element concentrations with increased deposition, which was observed for the RBF-SVR, Lin-SVR, and MLP models for Se and the Lin-SVR model for S.

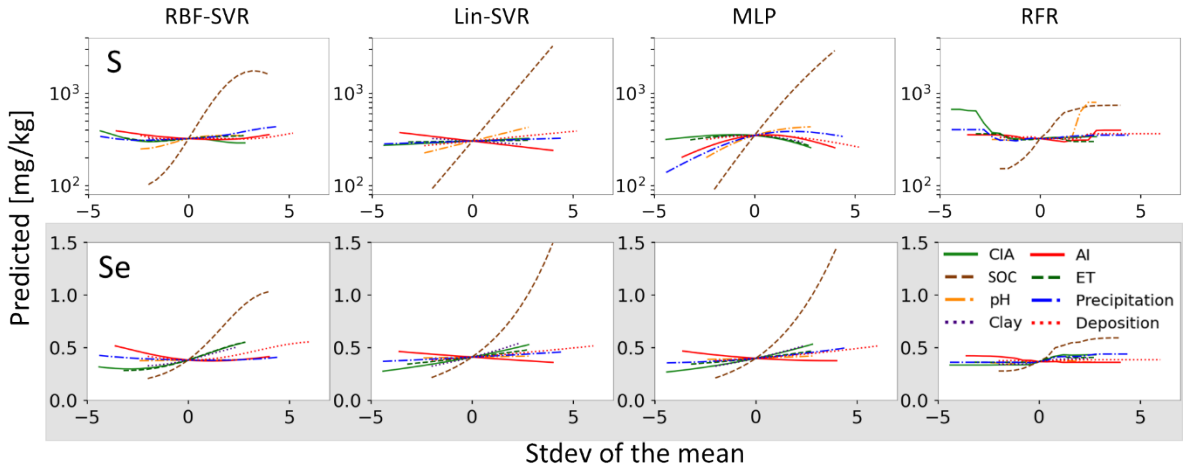

Fig. S8. machine-learning (ML) model sensitivity analyses for soil S and Se (gray box). All data are based on average values (n = 100). Panels in each column correspond to their heading. Abbreviations include sulfur (S), selenium (Se), radial basis function kernel for support vector regression (RBF-SVR), linear kernel for support vector regression (Lin-SVR), multilayer perceptron (MLP), and random forest regression (RFR), chemical index of alteration (CIA), soil organic carbon (SOC), aridity index (AI), evapotranspiration (ET), precipitation (Precip), and deposition (Dep).

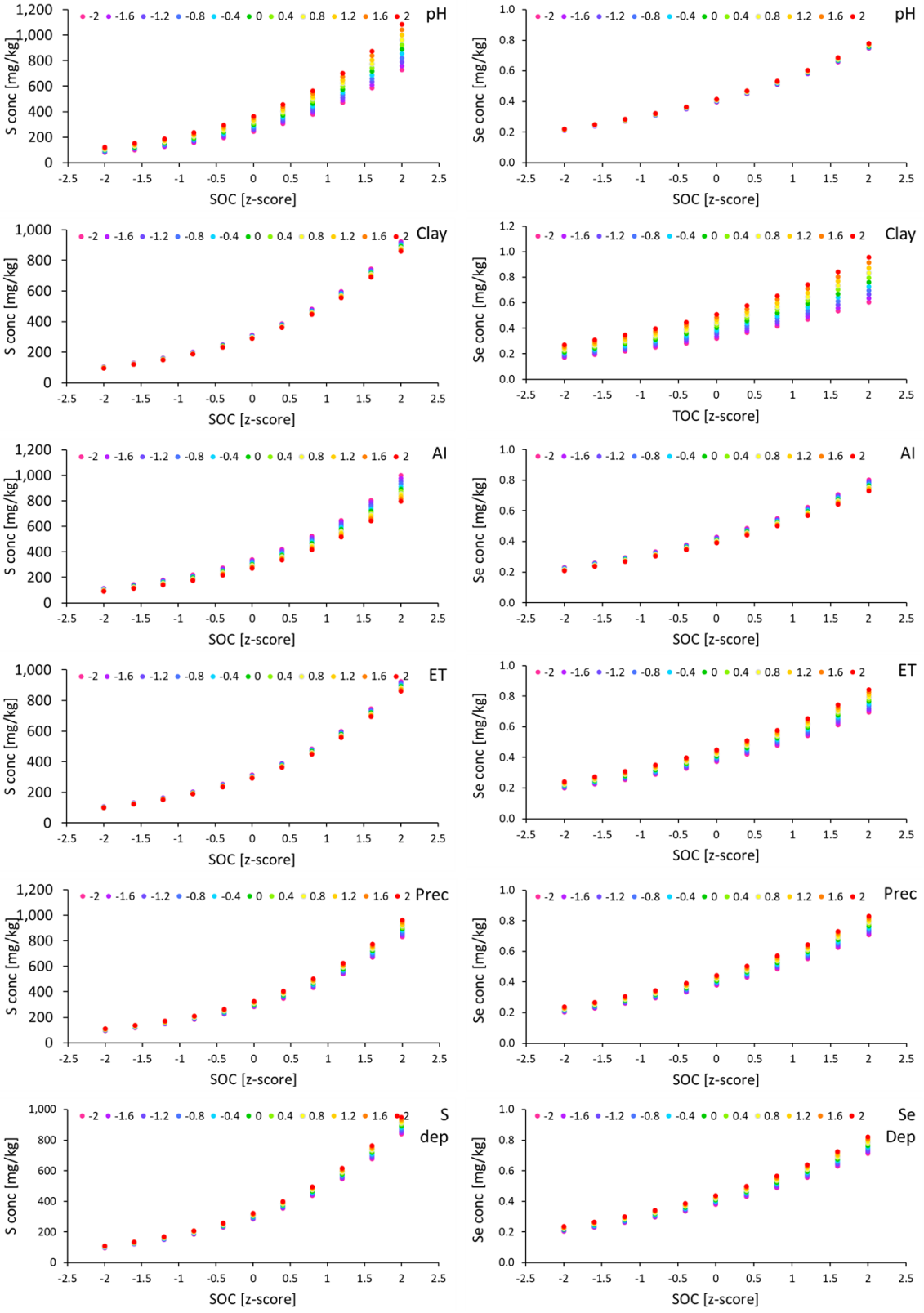

Fig. S9. Bivariate sensitivity analyses between SOC and all other predictor variables for S (left panels) and Se (right panels).

Finally, the element responses to changes in predictive variables were highly irregular in the RFR model. For example, elements were invariant to predictive variable changes across their domain, particularly for Se. The lack of a consistent relationship between the predictive variables and either element suggests that the RFR model developed highly non-linear and nuanced relationships between elements and variables, which further suggests that the RFR model is inconsistent with the expected mechanistic pattern. Overall, while all models have similar performance, the Lin-SVR models for both elements and the MLP model for Se are most mechanistically consistent with our expectations.

Future changes

For each model iteration, ML models were used to predict future changes in trace element concentrations based on projected changes in deposition, climate, and SOC throughout the 21st century. In general, there was some agreement between the four models. For example, S and Se concentrations in northern Europe were predicted to increase while concentrations in southern Europe were predicted to decrease; however, there are distinct differences in the maps, which are explained by some of the anomalous behavior observed in the sensitivity analyses. For example, we expect largely monotonic relationships between predictive variables, namely AI, and element concentrations; however, for the RFR model for S and the RBF-SVR models for S and Se, the models predicted increases in element concentrations at either AI extreme (Fig. S7). This explains why element concentrations are increasing in southern Spain despite our expectation that concentrations should drop with increasing aridity in the region (Fig. S8). Particularly for the RFR models, the predicted concentrations in the sensitivity analysis are irregular and nuanced (Fig. S7), which likely explains the lack of any broad-scale pattern and the highly variable predictions in the future data (Fig. S8). Qualitatively, we expect smooth transitions in element concentrations across

the continent, like shifts in broad scale changes in precipitation, SOC, and deposition. While our data aggregation on a 111 km scale can create discontinuities, the RFR model, particularly for Se, is the most discontinuous. Based on evidence of overfitting (Table S2), poor fitting (Fig. S5), and mechanistic discontinuities (Fig. S7 & S8), we argue that the RBF-SVR, MLP, and RFR models are least appropriate for S, and the RBF-SVR and RFR models are least appropriate for Se. The Lin-SVR models for S and Se and the MLP model for Se match our mechanistic expectation best of all the models. Furthermore, these maps best represent our expectation of relatively smooth changes in change across continental scales.

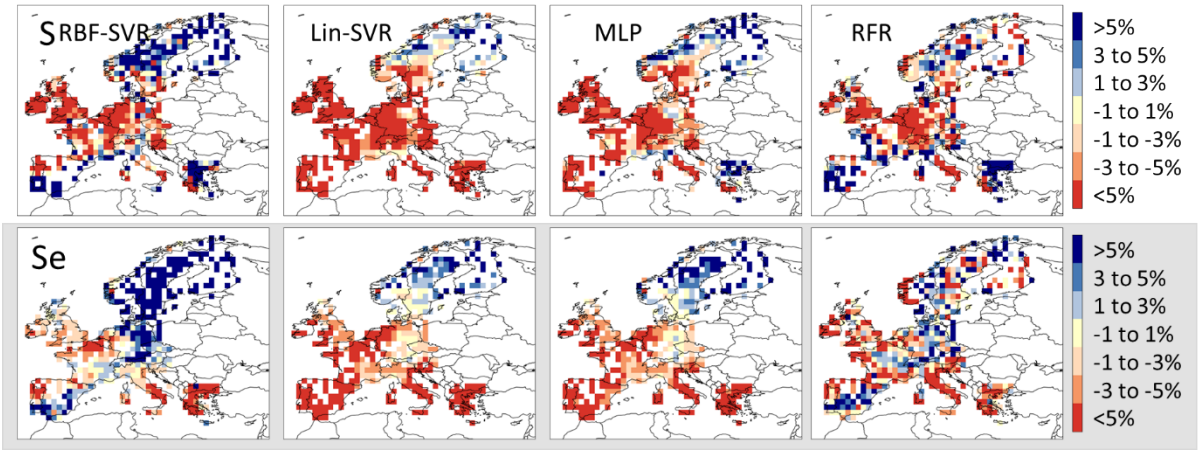

Fig. S10. Geographic distribution projected changes (i.e., (future modeled – current modeled)/current modeled) in Soil S and Se (gray box) by 2100 (c). Abbreviations include sulfur (S), selenium (Se), radial basis function support vector regression (RBF-SVR), linear support vector regression (Lin-SVR), multilayer perceptron (MLP), and random forest regression (RFR). Pixels that exceed accuracy and precision thresholds (all panels) or with missing data (c) were excluded. Projected changes are a result of changes in climate, deposition, and soil organic carbon. The pixel resolution of each panel is 111 km. Climate scenarios include A1FI for SOC, RCP 8.5 for climate, and SSP5–8.5 for S and Se emissions.

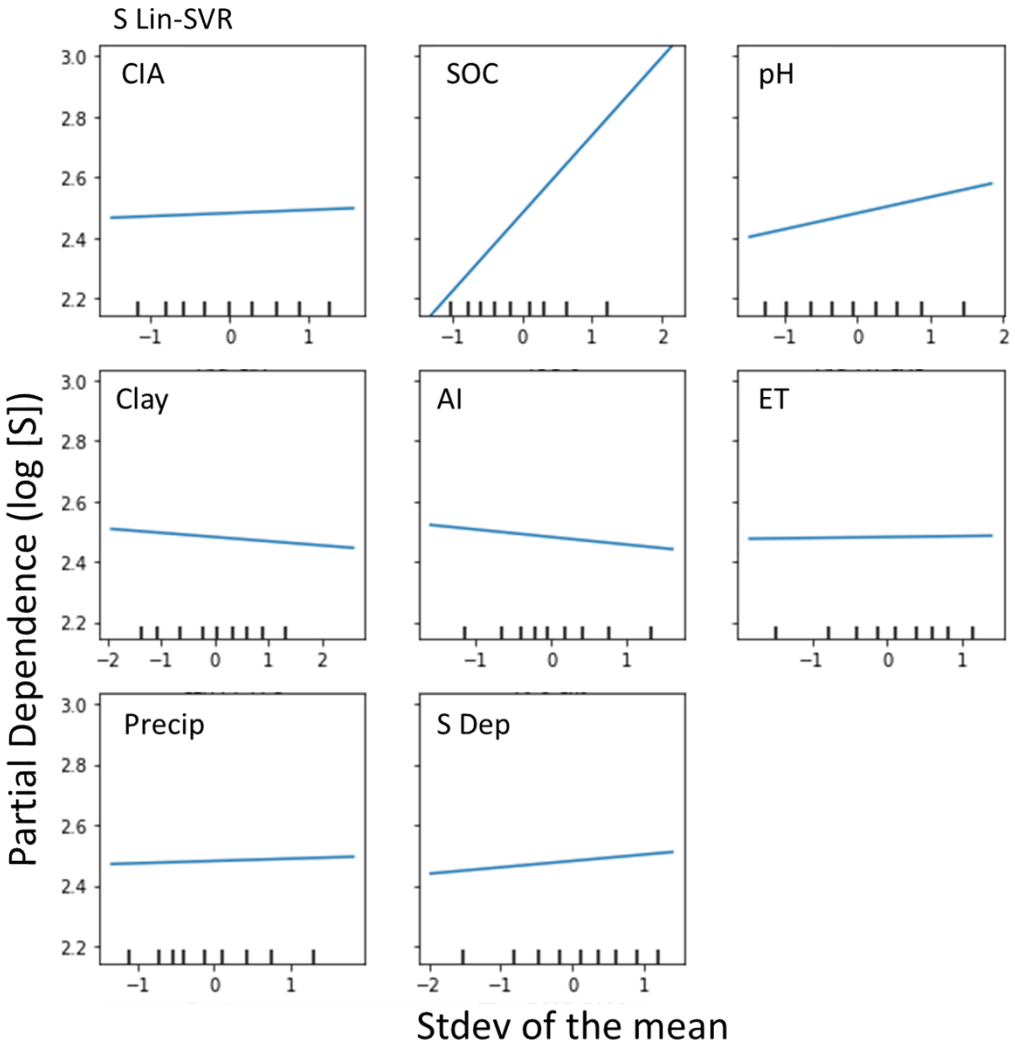

Fig. S11. Partial dependence plots for the Sulfur (S) linear-support vector regression (Lin-SVR) model. Data normalized center? We just need to better descriebe this graph.

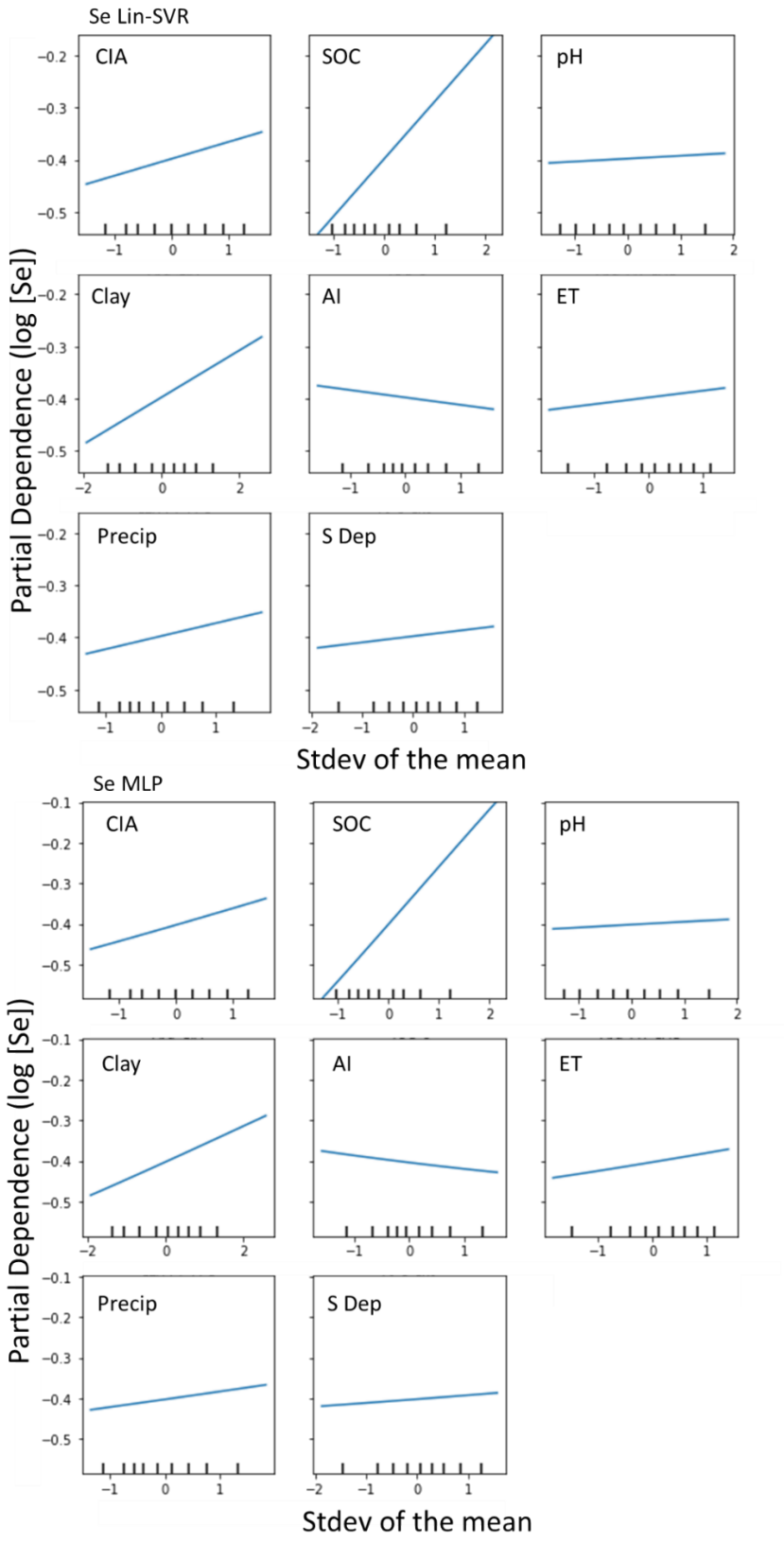

Fig. S12. Partial dependence plots for the Se linear- support vector regression (Lin-SVR) model (upper panels) and MLP (lower panels). See FigS11 comment

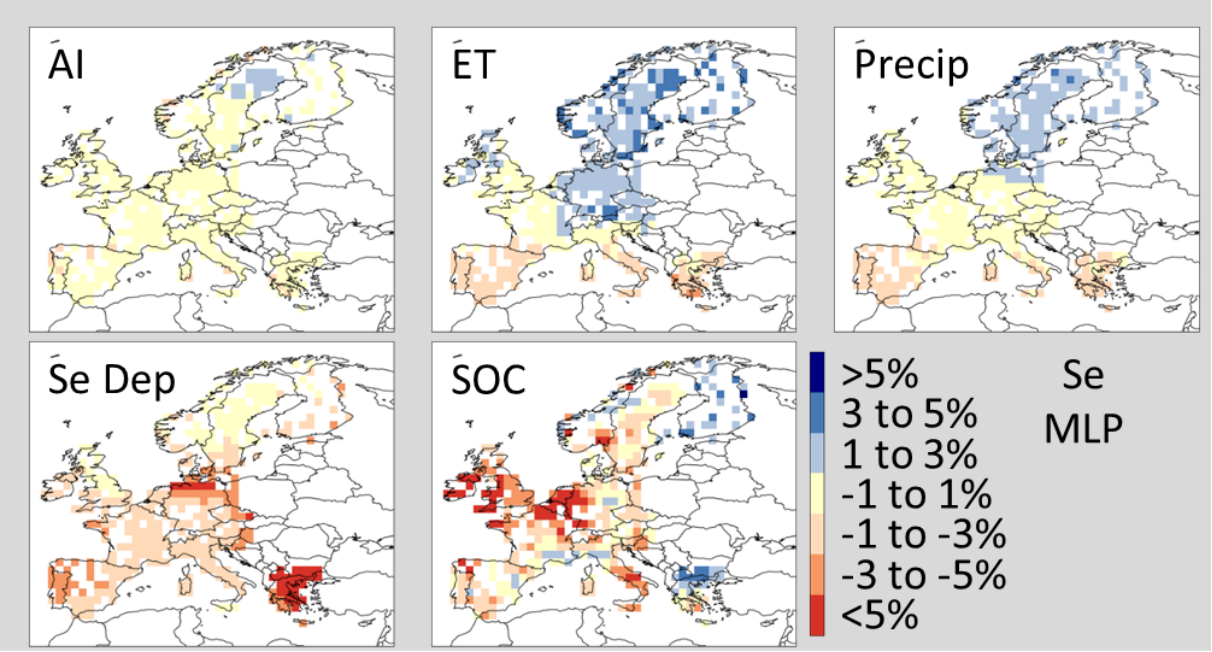

Fig. S13. Predicted percent change in soil Se concentrations by 2100 assuming an extreme climate change scenario (A1FI for SOC, RCP 8.5 for climate, and SSP5–8.5 for S and Se deposition). The independent effect of individual predictive variables are illustrated in each panel. In each panel, all other variables were modeled at their current values. Abbreviations include aridity index (AI), evapotranspiration (ET), precipitation (Precip), deposition (S or Se Dep), and soil organic carbon (SOC). Only results from the multilayer perceptron (MLP) for Se are illustrated. Results for linear support vector regression (Lin-SVR) are illustrated in Fig. 3. The pixel resolution of each panel is 111 km.

SUPPLEMENTARY REFERENCES

1. M. A. Lefsky, *Geophys. Res. Lett.*, 2010, **37**, L15401, DOI: 10.1029/2010gl043622.

2. P. D. Broxton, X. Zeng, W. Scheftic and P. A. Troch, *Journal of Applied Meteorology and Climatology*, 2014, **53**, 1996–2004, DOI: 10.1175/jamc-d-13-0356.1.

3. T. Hengl, J. Mendes de Jesus, G. B. M. Heuvelink, M. Ruiperez Gonzalez, M. Kilibarda, A. Blagotić, W. Shangguan, M. N. Wright, X. Geng, B. Bauer-Marschallinger, M. A. Guevara, R. Vargas, R. A. MacMillan, N. H. Batjes, J. G. B. Leenaars, E. Ribeiro, I. Wheeler, S. Mantel and B. Kempen, *PLoS One*, 2017, **12**, e0169748, DOI: 10.1371/journal.pone.0169748.

4. J. Smith, P. Smith, M. Wattenbach, S. Zaehle, R. Hiederer, R. J. A. Jones, L. Montanarella, M. D. A. Rounsevell, I. Reginster and F. Ewert, *Global Change Biol.*, 2005, **11**, 2141–2152, DOI: 10.1111/j.1365-2486.2005.001075.x.

5. D. Jacob, J. Petersen, B. Eggert, A. Alias, O. B. Christensen, L. M. Bouwer, A. Braun, A. Colette, M. Déqué, G. Georgievski, E. Georgopoulou, A. Gobiet, L. Menut, G. Nikulin, A. Haensler, N. Hempelmann, C. Jones, K. Keuler, S. Kovats, N. Kröner, S. Kotlarski, A. Kriegsmann, E. Martin, E. van Meijgaard, C. Moseley, S. Pfeifer, S. Preuschmann, C. Radermacher, K. Radtke, D. Rechid, M. Rounsevell, P. Samuelsson, S. Somot, J.-F. Soussana, C. Teichmann, R. Valentini, R. Vautard, B. Weber and P. Yiou, *Regional Environmental Change*, 2013, **14**, 563–578, DOI: 10.1007/s10113-013-0499-2.

6. J. Hartmann and N. Moosdorf, *Geochem. Geophys. Geosyst.*, 2012, **13**, Q12004, DOI: 10.1029/2012GC004370.

7. Center for International Earth Science Information Network (CIESIN) and Columbia University, Gridded population of the world, version 4 (GPWv4): Population density, http://dx.doi.org/10.7927/H4NP22DQ, (accessed 28 January 2017, DOI: 10.7927/H4NP22DQ).

8. NOAA National Geophysical Data Center, ETOPO1 1 arc-minute global relief model, https://doi.org/10.7289/V5C8276M, (accessed 12 January 2017, DOI: 10.7289/V5C8276M).

9. C. Reimann, M. Birke, A. Demetriades, P. Filzmoser and P. O'Connor, *Distribution of elements/parameters in agricultural and grazing land soil in Europe. Chemistry of Europe's agricultural soils. Part A: methodology and interpretation of the GEMAS data set*, Bundesanstalt für Geowissenschaften und Rohstoffe, Hannover, Germany, 2014.

10. A. Feinberg, A. Stenke, T. Peter, E.-L. S. Hinckley, C. T. Driscoll and L. H. E. Winkel, *Communications Earth & Environment*, 2021, **2**, 101, DOI: 10.1038/s43247-021-00172-0.

11. B. McCune, J. Grace and D. Urban, *Analysis of ecological communities*, MJM Software, Gleneden Beach, Oregon, US, 2002.

12. N. Yang, Z. Zheng and T. Wang, presented in part at the Proceedings of the 2019 11th International Conference on Machine Learning and Computing - ICMLC '19, 2019, DOI: 10.1145/3318299.3318367.

13. D. Chicco, *BioData Min.*, 2017, **10**, 35, DOI: 10.1186/s13040-017-0155-3.

14. A. W. Western, R. B. Grayson and G. Bloschl, *Annu. Rev. Earth Planet. Sci.*, 2002, **30**, 149–180, DOI: 10.1146/annurev.earth.30.091201.140434.

15. W. L. Silver, A. E. Lugo and M. Keller, *Biogeochemistry*, 1999, **44**, 301–328, DOI: 10.1007/Bf00996995.

16. V. K. Sharma, T. J. McDonald, M. Sohn, G. A. K. Anquandah, M. Pettine and R. Zboril, *Environ. Chem. Lett.*, 2014, **13**, 49–58, DOI: 10.1007/s10311-014-0487-x.

17. P. J. Edwards, *Sulfur cycling, retention, and mobility in soils: A review*, USDA Forest Service, Northeastern Research Station, Radnor, PA, US, 1998.

18. J. Tolu, Y. Thiry, M. Bueno, C. Jolivet, M. Potin-Gautier and I. Le Hecho, *Sci. Total Environ.*, 2014, **479–480**, 93–101, DOI: 10.1016/j.scitotenv.2014.01.079.

19. T. Delfosse, P. Delmelle and B. Delvaux, *Geoderma*, 2006, **136**, 716–722, DOI: 10.1016/j.geoderma.2006.05.009.

20. T. A. Sokolova and S. A. Alekseeva, *Eurasian Soil Sci.*, 2008, **41**, 140–148, DOI: 10.1134/S106422930802004X.

21. A. Pedescoll, R. Sidrach-Cardona, J. C. Sánchez and E. Bécares, *Ecol. Eng.*, 2013, **58**, 335–343, DOI: 10.1016/j.ecoleng.2013.07.007.

22. H. Renkema, A. Koopmans, L. Kersbergen, J. Kikkert, B. Hale and E. Berkelaar, *Plant Soil*, 2012, **354**, 239–250, DOI: 10.1007/s11104-011-1069-3.

23. M. D. Bertness and R. M. Callaway, *Trends Ecol. Evol.*, 1994, **9**, 191–193.

24. F. M. Padilla and F. I. Pugnaire, *Front. Ecol. Environ.*, 2006, **4**, 196–202, DOI: 10.1890/1540-9295(2006)004[0196:TRONPI]2.0.CO;2.

25. D. Subramanian, R. Subha and A. K. Murugesan, *Ecol. Indicators*, 2022, **135**, 108522, DOI: 10.1016/j.ecolind.2021.108522.

26. S. Zhou, A. P. Williams, B. R. Lintner, A. M. Berg, Y. Zhang, T. F. Keenan, B. I. Cook, S. Hagemann, S. I. Seneviratne and P. Gentine, *Nat. Clim. Change*, 2021, **11**, 38–44, DOI: 10.1038/s41558-020-00945-z.

27. H. Wen, J. Perdrial, B. W. Abbott, S. Bernal, R. Dupas, S. E. Godsey, A. Harpold, D. Rizzo, K. Underwood, T. Adler, G. Sterle and L. Li, *Hydrol. Earth Syst. Sci.*, 2020, **24**, 945–966, DOI: 10.5194/hess-24-945-2020.

28. F. Ayo and R. O. M, *J. Ecol. Nat. Environ.*, 2014, **6**, 182–189, DOI: 10.5897/jene2013.0433.

29. B. W. Stewart, R. C. Capo and O. A. Chadwick, *Geochim. Cosmochim. Acta*, 2001, **65**, 1087–1099, DOI: 10.1016/S0016-7037(00)00614-1.

30. L. K. McDonough, I. R. Santos, M. S. Andersen, D. M. O'Carroll, H. Rutlidge, K. Meredith, P. Oudone, J. Bridgeman, D. C. Gooddy, J. P. R. Sorensen, D. J. Lapworth, A. M. MacDonald, J. Ward and A. Baker, *Nat. Commun.*, 2020, **11**, 1279, DOI: 10.1038/s41467-020-14946-1.

31. G. D. Jones, B. Droz, P. Greve, P. Gottschalk, D. Poffet, S. P. McGrath, S. I. Seneviratne, P. Smith and L. H. E. Winkel, *Proc. Natl. Acad. Sci. U.S.A.*, 2017, **114**, 2848–2853, DOI: 10.1073/pnas.1611576114.

32. R. E. LaCroix, N. Walpen, M. Sander, M. M. Tfaily, J. L. Blanchard and M. Keiluweit, *Environ. Sci. Technol. Let.*, 2020, **8**, 92–97, DOI: 10.1021/acs.estlett.0c00748.