Extended information

Machine learning-based identification of key biotic and abiotic drivers of mineral weathering rate in a complex enhanced weathering experiment

Iris Janssens[1], Thomas Servotte[2], Tullia Calogiuri[3], Steven Mortier[1], Harun Niron[4], Thomas Corbett[5], Reinaldy P. Poetra[6], Lukas Rieder[6], Michiel Van Tendeloo[4], Abhijeet Singh[5], Steven Latré[1], Siegfried E. Vlaminck[4], Jens Hartmann[6], Jan Willem van Groenigen[3], Anna Neubeck[5], Alix Vidal[3], Ivan A. Janssens[7], Mathilde Hagens[8], Sara Vicca[4], and Tim Verdonck[2]

[1] *University of Antwerp - imec - IDLab, Department of Computer Science, Sint-Pietersvliet 7, 2000 Antwerp, Belgium*

[2] *University of Antwerp - imec, Department of Mathematics, Middelheimlaan 1, 2020 Antwerp, Belgium*

[3] *Wageningen University & Research, Soil Biology Group, Droevendaalsesteeg 3, 6708 PB, Wageningen, the Netherlands*

[4] *University of Antwerp, Department of Bioscience Engineering, Biobased Sustainability Engineering (SUSTAIN), Groenenborgerlaan 171, 2000 Antwerp, Belgium*

[5] *Uppsala University, Department of Earth Sciences, Villaväagen 16, 752 36 Uppsala, Sweden*

[6] *University of Hamburg, Institute for Geology, Center for Earth System Research and Sustainability, Bundesstraße 55, 20146 Hamburg, Germany*

[5] *Uppsala University, Department of Earth Sciences, Villaväagen 16, 752 36 Uppsala, Sweden*

[7] *University of Antwerp, department of Biology, Universiteitsplein 1, 2160 Wilrijk, Belgium*

[8] *Wageningen University & Research, Department of Soil chemistry, Droevendaalsesteeg 3, 6708 PB Wageningen, the Netherlands*

A   Data processing

A.1   Calculation of change in amount of dissolved inorganic carbon, solid inorganic carbon, total inorganic carbon and total alkalinity

This section outlines the methodology used to calculate the changes in the amount of dissolved inorganic carbon (DIC), solid inorganic carbon (SIC), total inorganic carbon (TIC), and total alkalinity (TA). Detailed equations and relevant measurements are provided to ensure clarity and reproducibility.

A.1.1   Change in dissolved inorganic carbon

The change in amount of DIC (ΔDIC) is defined as the difference between the final (DIC_f) and the initial DIC present at the beginning of the experiment (DIC_0):

$$\Delta DIC \, [mmol] = DIC_f [mmol] - DIC_0 \, [mmol] \tag{A.1}$$

	Solid bed	V_{irr} [l]	$DIC\left[\frac{mg}{l}\right]$	$TA\left[\frac{\mu mol}{l}\right]$
B_1	400g sand	5.6	11.875	1434.9
B_2	400g sand	5.6	11.088	1354.5
B_3	400g sand	8.4	11.783	1427.3
$B_{straw,1}$	400g sand + 10g straw	5.6	35.169	1981.1
$B_{straw,2}$	400g sand + 10g straw	8.4	13.589	1632.9
$B_{digestate,1}$	400g sand + 10g digestate	5.6	10.502	1445.4
$B_{alfalfa,1}$	400g sand + 10g alfalfa mixture	5.6	26.387	2304.8
$B_{biochar,1}$	400g sand + 20g biochar	5.6	15.452	1761.3
$B_{biochar,2}$	400g sand + 40g biochar	5.6	18.387	2053.5
$B_{biochar,3}$	400g sand + 20g biochar	8.4	14.708	1698.6

Table A1: Concentrations of dissolved inorganic carbon (DIC) and total alkalinity (TA), measured in leached water of blank columns, with a solid bed consisting only of sand, biochar and organic amendments.

Here, DIC_f is calculated using the DIC concentration measured after 8 weeks (DIC_{conc}) and the irrigated volume (V_{irr}):

$$DIC_f\,[mmol] = DIC_{conc}\left[\frac{mg}{l}\right] \cdot V_{irr}\,[l] \cdot \frac{1\,[mmol]}{12.001\,[mg]} \tag{A.2}$$

The DIC_0 is the sum of the DIC present in the irrigated water and the DIC stemming from the organic amendments and biochar. To calculate the DIC stemming from the irrigated water, the DIC concentration of columns with a solid bed consisting only of 400g sand (coined B), was measured (DIC_{conc}^{B}) and averaged (Tab. A1).

$$DIC_{conc}^{water}\left[\frac{mg}{l}\right] = \overline{DIC_{conc}^{B_i}\left[\frac{mg}{l}\right]} \tag{A.3}$$

The DIC stemming from the organic amendments and biochar was based on DIC concentration measurements of columns with a solid bed made out of only sand and straw (B_{straw}), digestate ($B_{digestate}$), alfalfa-dominated mixture ($B_{alfalfa}$) or biochar ($B_{biochar}$) (Tab. A1):

$$DIC_{conc}^{matter}\left[\frac{mg}{g\,matter}\right] = \frac{\overline{(DIC_{conc}^{B_{matter,i}}[\frac{mg}{l}] - DIC_{conc}^{water}[\frac{mg}{l}]) \cdot V_{irr}^{B_{matter,i}}\,[l]}}{m_{org,i}^{B_{matter,i}}\,[g]} \tag{A.4}$$

Here, *matter* can be either straw, digestate, the alfalfa mixture and biochar.
Finally, the initial DIC was calculated using:

$$DIC_0\,[mmol] = \left(DIC_{conc}^{water}\,[mg] \cdot V_{irr}\,[l] + \sum_{matter} DIC_{conc}^{matter}\left[\frac{mg}{g\,matter}\right] \cdot m_{matter}\,[g]\right) \cdot \frac{1\,[mmol]}{12.001\,[mg]} \tag{A.5}$$

Here, the summation goes over straw, digestate, the alfalfa mixture and biochar.

A.1.2 Change in amount of solid inorganic carbon

Analogously to ΔDIC, change in amount of SIC (ΔSIC) is the difference between the final (SIC_f) and the initial SIC that is added to the system through the input materials (SIC_0):

$$\Delta SIC\,[mmol] = SIC_f\,[mmol] - SIC_0\,[mmol] \tag{A.6}$$

Before measuring the SIC concentration (SIC_{conc}) of the solid samples (using *C analysis, Flash 2000 CN Soil Analyser, Interscience, Antwerp, Belgium*), loss of ignition (LOI) is performed, which lead to mass loss. Therefore, SIC_f is calculated using:

$$SIC_f\,[mmol] = 10 \cdot SIC_{conc}\,[mass\%] \cdot LOI\,correction \cdot m_{solid}\,[g] \cdot \frac{1\,[mmol]}{12.001\,[mg]} \tag{A.7}$$

2

where m_{solid} is the mass of the added minerals, organic amendments and biochar; and the *LOI correction* factor accounts for sample mass loss during LOI:

$$LOI\ correction = 1 - \frac{m_{solid\ sample}^{post-LOI}\ [g] - m_{crucible}\ [g]}{m_{solid\ sample}^{pre-LOI}\ [g] - m_{crucible}\ [g]} \tag{A.8}$$

with $m_{solid\ sample}^{pre-LOI}$ and $m_{solid\ sample}^{post-LOI}$ the gross sample mass before and after LOI, and $m_{crucible}$ the mass of the crucible holding the solid sample.

SIC_0 is the SIC present at the start of the experiment and is calculated based on SIC measurements of the added minerals, organic amendments and biochar (Tab. A2) and their masses in the column ($m_{sample,X}$):

$$SIC_0\ [mmol] = 10 \cdot \sum_X SIC_{conc}^X\ [mass\%] \cdot LOI\ correction_X \cdot m_{sample,X}\ [g] \cdot \frac{1\ [mmol]}{12.001\ [mg]} \tag{A.9}$$

where the summation goes over the added minerals, organic amendments and biochar; and the *LOI correction* factor again accounts for sample mass loss during LOI (Eq. A.8).

Matter		SSA $[m^2/g]$	$m_{sample,X}^{pre-LOI}$ $[g]$	$m_{sample,X}^{post-LOI}$ $[g]$	$m_{crucible,X}$ $[g]$	SIC_{conc}^X $[mass\%]$
Lava	$\leq$ 0.063 mm	1.896	37.5	37.4	20.9	0.009
	0.063-2 mm	0.513	41.9	41.9	20.2	0.006
Basalt	$\leq$ 0.063 mm	10.577	38.4	38.3	19.7	0.134
	0.063-3 mm	6.972	40.2	40.0	21.3	0.068
Dunite	$\leq$ 0.063 mm	7.050	36.2	35.8	18.8	0.031
	0.063-0.125 mm	0.959	45.2	45.1	19.8	0.005
	0.125-1 mm	1.134	51.2	51.1	20.0	0.008
Steelslag	$\leq$ 0.1 mm	13.246	32.6	31.8	18.8	1.922
	0.1-0.5 mm	9.154	39.9	38.9	19.3	1.858
	0.5-3 mm	4.747	38.8	38.3	18.8	1.359
	mixed ($\leq$ 3 mm)	10.271	29.6	29.2	19.0	0.992
Digestate		-	22.2	20.2	20.0	2.272
Straw		-	21.0	20.0	20.0	4.957
Alfalfa mixture		-	23.7	23.3	23.3	3.339
Biochar		-	23.8	20.5	19.5	0.263

SSA = Specific surface area, LOI = loss of ignition, SIC = solid inorganic carbon. SSA measurements of minerals are obtained using Brunauer-Emmett-Teller method. $m_{solid\ sample}^{pre-LOI}$ and $m_{solid\ sample}^{post-LOI}$ are the gross sample masses before and after LOI, $m_{crucible}$ is the mass of the crucible holding the sample and SIC is the measured solid inorganic carbon concentration.

Table A2: Specific surface area measurements of minerals, loss of ignition and solid inorganic carbon measurements of samples of minerals, biochar and organic amendments.

A.1.3   Change in total inorganic carbon

In this study, total inorganic carbon is defined as the sum of solid inorganic carbon and dissolved inorganic carbon. Therefore, the final TIC, initial TIC and change in TIC are calculated as:

$$TIC_f\ [mmol] = SIC_f\ [mmol] + DIC_f\ [mmol] \tag{A.10}$$

$$TIC_0\ [mmol] = SIC_0\ [mmol] + DIC_0\ [mmol] \tag{A.11}$$

$$\Delta TIC\ [mmol] = \Delta SIC\ [mmol] + \Delta DIC\ [mmol] \tag{A.12}$$

A.1.4 Change in total alkalinity

The change in total alkalinity (ΔTA) is the difference between the final (TA_f) and the initial TA that results from non-weathering processes (TA_0):

$$\Delta TA\,[mmol] = TA_f\,[mmol] - TA_0\,[mmol] \tag{A.13}$$

The final TA is calculated from the measured TA concentration at the end of the experiment (TA_{conc}) and the irrigated volume:

$$TA_f\,[mmol] = 10^{-3}\,TA_{conc}\left[\frac{\mu mol}{l}\right] \cdot V_{irr}\,[l] \tag{A.14}$$

The initial TA is calculated analogously to the initial DIC (Equations A.3-A.5), i.e., based on TA concentration measurements of batches with a solid bed consisting only of 400g sand (B) and batches with a solid bed made out of only organic amendments and biochar (B_{straw}, $B_{digestate}$, $B_{alfalfa}$, $B_{biochar}$) (Tab. A1):

$$TA_{conc}^{water}\left[\frac{mmol}{l}\right] = 10^{-3}\,\overline{TA_{conc}^{B_i}\left[\frac{\mu mol}{l}\right]} \tag{A.15}$$

$$TA_{conc}^{matter}\left[\frac{mmol}{g\,matter}\right] = 10^{-3}\,\overline{\frac{(TA_{conc}^{B_{matter,i}}[\frac{\mu mol}{l}] - TA_{conc}^{water}[\frac{\mu mol}{l}]) \cdot V_{irr}^{B_{matter,i}}\,[l]}{m_{matter,i}^{B_{matter,i}}\,[g]}} \tag{A.16}$$

$$TA_0\,[mmol] = TA_{conc}^{water}\left[\frac{mmol}{l}\right] \cdot V_{irr}\,[l] + \sum_{matter} TA_{conc}^{matter}\left[\frac{mmol}{g\,matter}\right] \cdot m_{matter}\,[g] \tag{A.17}$$

where the summation goes over straw, digestate, the alfalfa mixture and biochar.

A.2 Transformation of model features

Some features were transformed before being used as input for the machine learning (ML) models. The relevant equations for such features are provided below.

The total surface area of mineral type M is calculated using:

$$SA_M\ [m^2] = \sum_{G}^{grain\ sizes} mass_{MG}\ [g] \cdot SSA_{MG}\ \left[\frac{m^2}{g}\right] \tag{A.18}$$

where $mass_{MG}$ is the added mass, and SSA_{MG} is the specific surface area of mineral type M with grain size G (Table A2).

The feature *chlorotica ratio* is defined as the ratio of initially live *A. chlorotica* earthworms with respect to the total number of initially live earthworms (i.e., the sum of the number of *A. chlorotica* earthworms and the number of *A. caliginosa* earthworms):

$$chlorotica\ ratio = \frac{number\ of\ initially\ live\ A.\ chlorotica\ earthworms}{number\ of\ initially\ live\ earthworms} \tag{A.19}$$

And analogously, the feature † *chlorotica ratio* is defined as:

$$\dagger\ chlorotica\ ratio = \frac{number\ of\ initially\ dead\ A.\ chlorotica\ earthworms}{number\ of\ initially\ dead\ earthworms} \tag{A.20}$$

A.3 Analysis of noise in data

To quantify the noise in the data, some of the input combinations of the columns was repeated. This resulted in 28 combinations that were repeated at least three times for ΔDIC, 30 for ΔSIC, 27 for change in amount of TIC (ΔTIC) and 27 for change in amount of TA (ΔTA). For each set of repeats, the standard deviation of the weathering indicators (ΔDIC, ΔSIC, ΔTIC and acDTA) was calculated and rescaled by dividing by the interquartile range of the indicator of the full dataset (Fig. A1). As steel slag generally lead to different scales of the indicators, this rescaling was done separately for the subset of columns with and without steel slag. For all indicators, the resulting rescaled standard deviation was relatively high, especially for ΔSIC and, to a lesser extent, ΔTIC (Fig. A1).
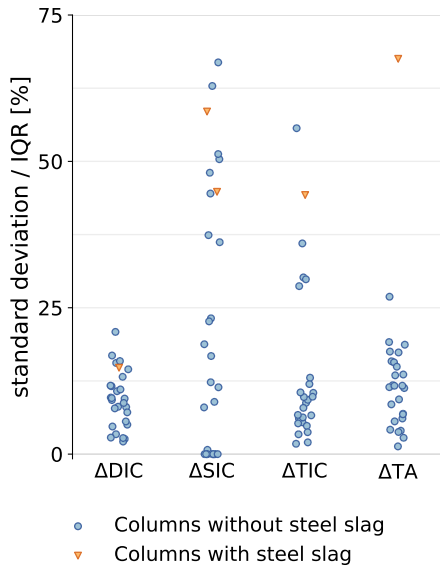

Figure A1: Quantification of the noise in data of the changes of dissolved inorganic carbon (ΔDIC), solid inorganic carbon (ΔSIC), total inorganic carbon (ΔTIC) and total alkalinity (ΔTA). The y axis depicts the standard deviations of columns that were repeated at least three times, rescaled by interquartile range (IQR) of the full dataset. The rescaling was done separately for the datasets with (orange) and without steel slag (blue).

To assess whether this noise resulted from certain inputs of the columns, we used a Random Forest (RF) to predict the standard deviations of the target values of repeated combinations for each of the indicators (ΔDIC, ΔSIC, ΔTIC, ΔTA). As the number of repeated columns was limited, leave one out cross validation (CV) was used. However, the model's performances were poor for all indicators, except for the prediction of ΔSIC (Tab. A3). A subsequent Shapley Additive exPlanations (SHAP) analysis revealed that high standard deviations were predicted for columns with steel slag (A2). Therefore, the prediction of ΔSIC was repeated for columns without steel slag, which resulted in poor performances (Tab. A3). This suggests that, while steel slag leads to noisy measurements for SIC, the remaining noise of the indicators likely arises from other factors than the column inputs, such as sampling inconsistencies from imperfectly homogenized columns or irregular irrigation.

		MAE	MSE	RMSE	R^2
ΔDIC	RF	0.37	0.45	0.20	-0.50
	naive	0.30	0.38	0.14	-0.08
ΔSIC	RF	3.2	8.4	70	0.73
	naive	8.2	16.5	272	-0.07
ΔTIC	RF	2.8	7.9	62	-0.03
	naive	3.5	8.1	65	-0.08
ΔTA	RF	1.6	5.1	26	-0.20
	naive	1.8	4.8	23	-0.08
ΔSIC	RF	0.97	1.2	1.4	0.23
(without steel slag)	naive	1.21	1.4	2.0	-0.08

Table A3: Performance of the Random Forest prediction of the standard deviation of ΔDIC, ΔSIC, ΔTIC and ΔTA of repeated measurements; and of ΔSIC of repeated measurements without steel slag. Depicted performance metrics are mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and determination coefficient (R^2). The performace of the naive regressor, which predicts the mean target value for the whole dataset, is added as a baseline.
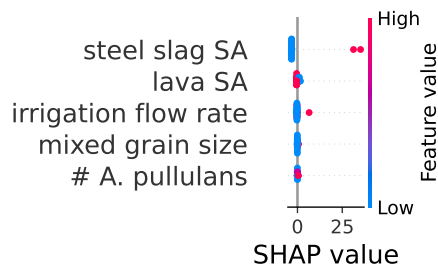

Figure A2: SHapley Additive exPlanation (SHAP) values of the top 5 most important features in the random forest prediction of the standard deviation of the change in solid inorganic carbon of repeated measurements.