Electronic Supplementary Material (ESI) for Environmental Science: Nano.

Influential parameters of surface waters on the formation of coating onto TiO_2 nanoparticles under natural conditions

Narjes Tayyebi Sabet Khomami [a], Parthvi Mayurbhai Patel [a], Cynthia Precious Jusi [a], Vanessa Trouillet [b], Jan David[a], Gabrielle. E. Schaumann[a], Allan Philippe [a*]

* Corresponding author

[a] iES Landau, Institute for Environmental Sciences, Koblenz-Landau University, Fortstrasse 7, 76829 Landau, Germany.

[b] Institute for Applied Materials (IAM) and Karlsruhe Nano Micro Facility (KNMF), Karlsruhe Institute of Technology (KIT), 76344 Eggenstein-Leopoldshafen, Germany.

Table S1: The locations of surface water (SW) sites and their descriptions.

Abbr.	Site	GPS Location	Type of landscape	Description
SW1	Rehbach	49° 21’ 20” N 8° 9’ 19” E	urban	Is tributary of the Speyerbach river which flows through the Winziger Wassergescheid in Neustadt Weinstrasse.
SW2	Speyerbach	49°19'04.8"N 8°26'49.5"E	urban	The Speyerbach is a left tributary of the Rhine river and flows through the southern palatinate forest as splits into smaller water courses before emptying out into the Rhine.
SW3	Bischofsweiher	49°20'40.4"N 8°05'18.2"E	forest	Bischofsweier is an artificial lake dammed from inflows from the Kaltenbrunnertalbach stream and serves as a recreational fishing lake.
SW4	Kaltenbrunnertalbach	49°20'40.4"N 8°05'18.2"E	forest	Kaltenbrunnertalbach is a stream that flows from the northern summit of Hüttenhohl and maintains its course through the southern palatinate forest before emptying into Rehbach.
SW5	Modenbach	49°16'12.4"N 8°10'58.4'' E	agricultural	Modenbach is a stream, just under 30 kilometers long, and a right-hand tributary of the Speyerbach.
SW6	Neuhofener Altrhein	49°25'41.7"N 8°27'18.3"E	agricultural/ urban	This lake is primarily reserved for nature conservation with the northern beach of the lake serving as a shallow bathing area open to the public while fishing is allowed in other sections. There is a constant danger of mass development of cyanobacteria.
SW7	Kiefweiher	49°26'25.2"N 8°28'13.6"E	agricultural/ urban	Kiefweiher differs from all other lakes in the city as it is directly connected to the Rhine. This has a special effect on Rhine flood plains when the Kiefweiher overflows its banks due to outflows form the Rhine river.
SW8	Rhein	49°19'08.2"N 8°26'59.8"E	urban	The Rhine is a major European river, with its source in Switzerland. Flowing through Germany’s Rhineland and the Netherlands to eventually empty into the North Sea.
SW9	Schwanenweiher	49°11'47.7"N 8°07'18.5"E	urban	It is an artificial pond located in the heart of Landau. The bank of the pond serves as a spot for recreational activities while the pond serves as a habitat for swans and fish. On Monday, July 1st, 2019, the temperature in pond increased to over 28 °C resulting in the reduction of biological oxygen demand driving most of the fish to the surface (www.pfalz-express.de).
SW10	Hainbach	49°14'09.9"N 8°04'25.3"E	Forest/agricultural	It is a tributary of the Speyerbach that runs through palatinate forest and serves as a natural irrigation for the vineyards and agricultural fields.

Table S2: The average of surface water parameters between the first and last day of exposure for spring and summer samples.[1]

Site	pH	Temp. (°C)	EC (µS/cm)	DOC (mg/L)	F^- (mg/L)	Cl^- (mg/L)	NO_3^- (mg/L)	SO_4^{2-} (mg/L)	PO_4^{3-} (mg/L)[2]	Na^+ (mg/L)	K^+ (mg/L)	Ca^{2+} (mg/L)	Mg^{2+} (mg/L)	Flu. fulvic/ humic acid[3]	UV 254 /210	Flu. Protein[4]
SW1-Sp.	7.9	11.2	147.5	2.3	0.04	10.6	4	14.4	0.2	9.1	4	14.9	4.1	0.9	0.03	1
SW1-Su.	7.8	13.8	154.0	8.8	0.01	6.0	1.9	7.3	0.2	8.4	3.4	15.7	3.3	0.6	0.04	0
SW2-Sp.	8	13.1	347.5	3.8	0.04	25.3	6.7	27.8	0.2	19.5	6.6	24.9	5.8	1.0	0.05	1
SW2-Su.	7.6	15.3	323.5	4.5	0.03	8.8	2.7	6.8	0.2	21.2	6.4	20.3	3.7	0.7	0.05	1
SW3-Sp.	7.4	11.1	122.0	3.0	0.05	9.2	4.3	14.2	<LOD	5.9	2.5	10.4	4.6	0.7	0.11	0
SW3-Su.	6.8	14.5	79.0	10.5	0.05	3.3	0.5	4.1	<LOD	2.3	1.2	5.3	1.1	0.5	0.29	0
SW4-Sp.	7.4	11.2	123.5	3.9	0.05	12.0	2.8	13.6	<LOD	5.7	2.7	9.4	4.0	0.6	0.10	0
SW4-Su.	6.8	14.6	84.0	6.2	0.07	3.1	0.4	3.5	<LOD	5	3	9.9	3.2	0.5	0.30	0
SW5-Sp.	8.1	11.3	357.0	3.1	0.07	11.6	5.2	21.6	<LOD	8.1	2.9	42.8	12.8	0.8	0.04	1
SW5-Su.	7.8	16.3	444.5	11.6	0.08	9.3	2.8	14.1	<LOD	6	1.8	42.8	9.4	0.5	0.05	0
SW6-Sp.	7.9	16.4	748.0	9.7	0.06	49.0	2.8	180	<LOD	24.3	6.7	86.5	21.3	0.7	0.07	1
SW6-Su.	7.9	21.6	817.0	13.6	0.08	20.1	0.3	40.4	<LOD	24.9	7	102.1	20.1	0.5	0.29	1
SW7-Sp.	8.4	16.2	338.0	3.5	0.07	20.3	4.3	27.6	<LOD	13.6	2.5	46.7	8.6	0.8	0.06	1
SW8-Sp.	8.2	13.6	349.5	2.9	0.07	19.6	5.4	25.8	<LOD	13.8	2.4	48.8	8.5	0.8	0.05	1
SW8-Su.	7.8	19.7	329.0	12.5	0.08	4.6	1.4	7.5	<LOD	10.5	2.1	34.5	6.9	0.5	0.05	0
SW9-Sp.	9.1	14.8	495.0	17.1	0.07	59.5	1.6	46.2	<LOD	38.9	10	30.5	11.9	0.9	0.19	1
SW9-Su.	8.6	23.4	852.5	19.3	0.09	133.2	1.8	30.8	0.2	62.3	15.5	63.1	18.8	0.6	0.17	1
SW10-Su.	7.5	16.4	119.5	5.6	0.07	4.1	1.7	6.6	0.1	3.5	2.4	16.2	3.8	1.0	0.15	0

[1] Sp. = spring (18.04.2019-22.04.2019) and Su. = summer (12.09.2019-16.09.2019). The data for SW10-Sp. and SW7-Su. are not available because the samples could not be retrieved after exposure.

[2] LOD (limit of detection) of phosphate = 0.08 mg/L.

[3] Fluorescence intensity ratio of fulvic/humic acid (fulvic: fluorescence intensity measured at Ex/Em ~ 340/430 and humic: fluorescence intensity measured at Ex/Em ~ 250/430).

[4] Tryptophane-like proteins in SWs were shown as “Flu. Protein” (on fluorescence map: Ex/Em ~ 270/330). Presence of “Fluo. Protein” = 1, Absence of “Flu. Protein” = 0.

Table S2-1: Calculations[1] of saturation index (SI) based on the parameters of Table S2.

Site	Ionic Strength	Saturation index.[2] (calcite)[3]
SW1-Sp.	2.06e-03	-0.944
SW1-Su.	1.74e-03	-1.083
SW2-Sp.	3.47e-03	-0.532
SW2-Su.	2.26e-03	-1.366
SW3-Sp.	1.58e-03	-2.089
SW3-Su.	5.74e-04	-3.512
SW4-Sp.	1.49e-03	-2.130
SW4-Su.	1.04e-03	-3.254
SW5-Sp.	4.38e-03	-0.124
SW5-Su.	3.64e-03	-0.659
SW6-Sp.	9.97e-03	-0.268
SW6-Su.	8.33e-03	-0.093
SW7-Sp.	4.95e-03	0.545
SW8-Sp.	4.72e-03	0.149
SW8-Su.	2.94e-03	-0.708
SW9-Sp.	1.03e-02	1.553
SW9-Su.	9.73e-03	1.076
SW10-Su.	1.53e-03	-1.638

[1] Calculations were performed using “visual MINTEQ 3.1” considering the atmospheric pressure of 0.00038 atm for CO_2.

[2] Saturation Index (SI) = log IAP – log K_{sp} (IAP: Ion activity product and K_{sp}: Solubility Product)
SI > 0: oversaturation (precipitation)
SI = 0: apparent equilibrium
SI < 0: undersaturation

[3] Calcite: $CaCO_3$

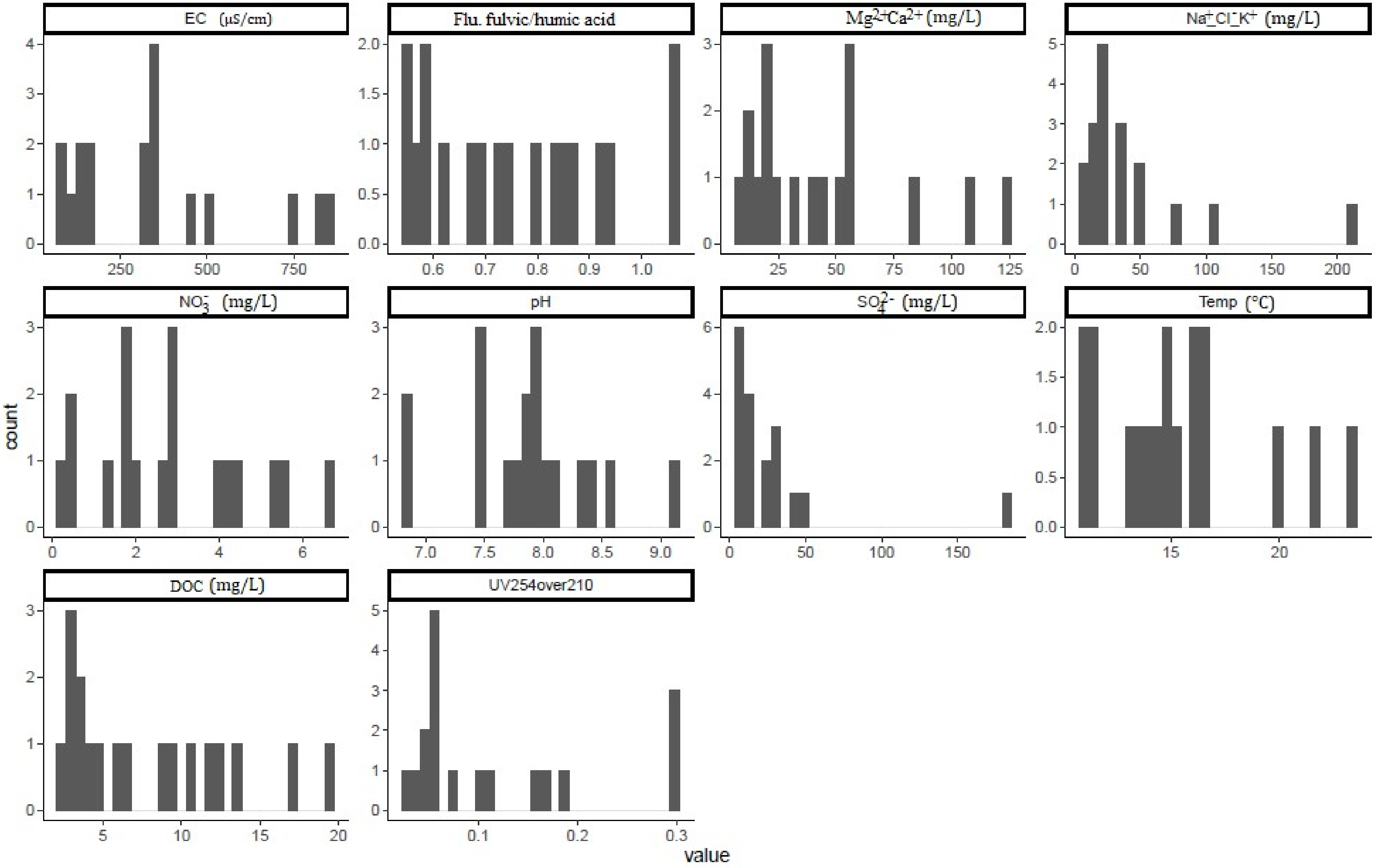

Figure S1: Density plots of the physiochemical parameters of the surface waters.

Quality control of the dialysis bags in different surface waters Using NMR relaxometry

Membrane fouling reduces the permeability of the dialysis bag due to pore obstruction by biotic or abiotic materials in surface waters. In our previous study, ^{1}H-NMR relaxometry was used as a simple and efficient method for in situ estimation of a pore size distribution averaged over the whole membrane; it was shown that the pore system of dialysis bags in surface water (Queich river) was not clogged after 7 days of exposure.[1] The same method was used here to investigate whether the dialysis bags remain efficient (the pores are not clogged) under exposing them to different surface waters with contrasting parameters. The T_2 distributions of dialysis membrane samples depict two distinguished T_2 peaks representing the hierarchical pore structure in dialysis bags (**Figure S2**). The larger T_2 (around 2000 – 3000 ms) distribution is the result of a mixed contribution between the water molecules absorbed on the surface of the membrane and free state while the smaller T_2 (150 – 200 ms) indicates water trapped in small pores. Clogged dialysis bags obtained after drying show a T_2 distribution mode around 70 ms.[1] Since all dialysis bags containing n-TiO_2 exposed to SWs during spring showed a T_2-distribution mode larger than 100 ms, we can assume that clogging did not occur in the tested surface waters.

Furthermore, Tukey test and box plots of T_2 (**Figure S3**) showed that the T_2-distributions modes of SW6 and SW9 varied significantly from each other (p-value = 0.012), but not with the other sites. Interestingly, SW6 and SW9, despite being relatively similar in terms of water parameters, show contrasting behaviours towards dialysis membrane. Hence, only SW with extreme composition could result in significant differences in the pore size distribution of the dialysis bags; however, this effect is not relevant for most of the SWs and is not expected to affect the composition of the natural coating.

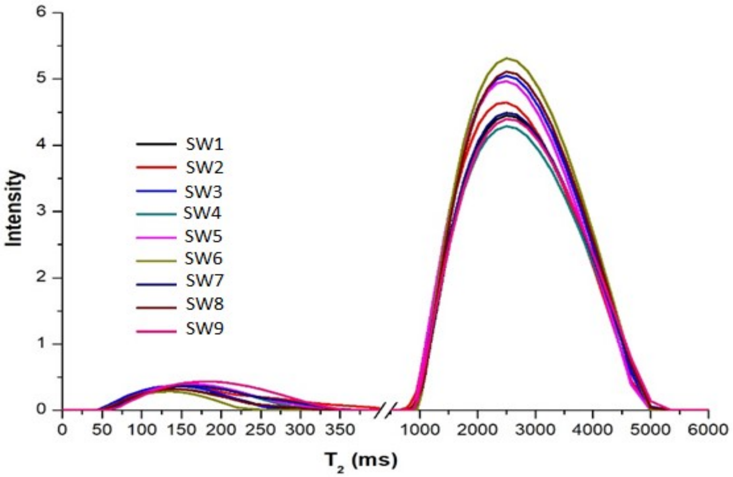

Figure S2: T_2 distribution comparison of dialysis bags containing n-TiO_2 exposed to surface waters SW1-SW9 during spring obtained using ^{1}H-NMR relaxometry. Sample SW10 was not retrieved.

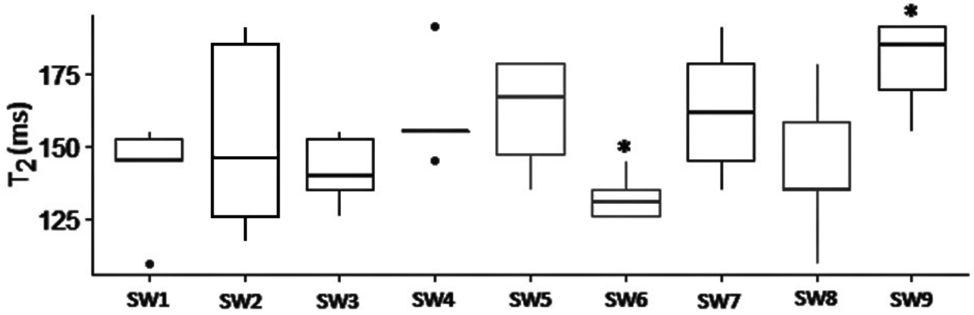

Figure S3: Box plots showing variances in T_2-distribution mode (small pores) within each six replicates in spring experiment. Sample SW10 was not retrieved. Asterisks show the samples with significant difference based on a Tukey test.

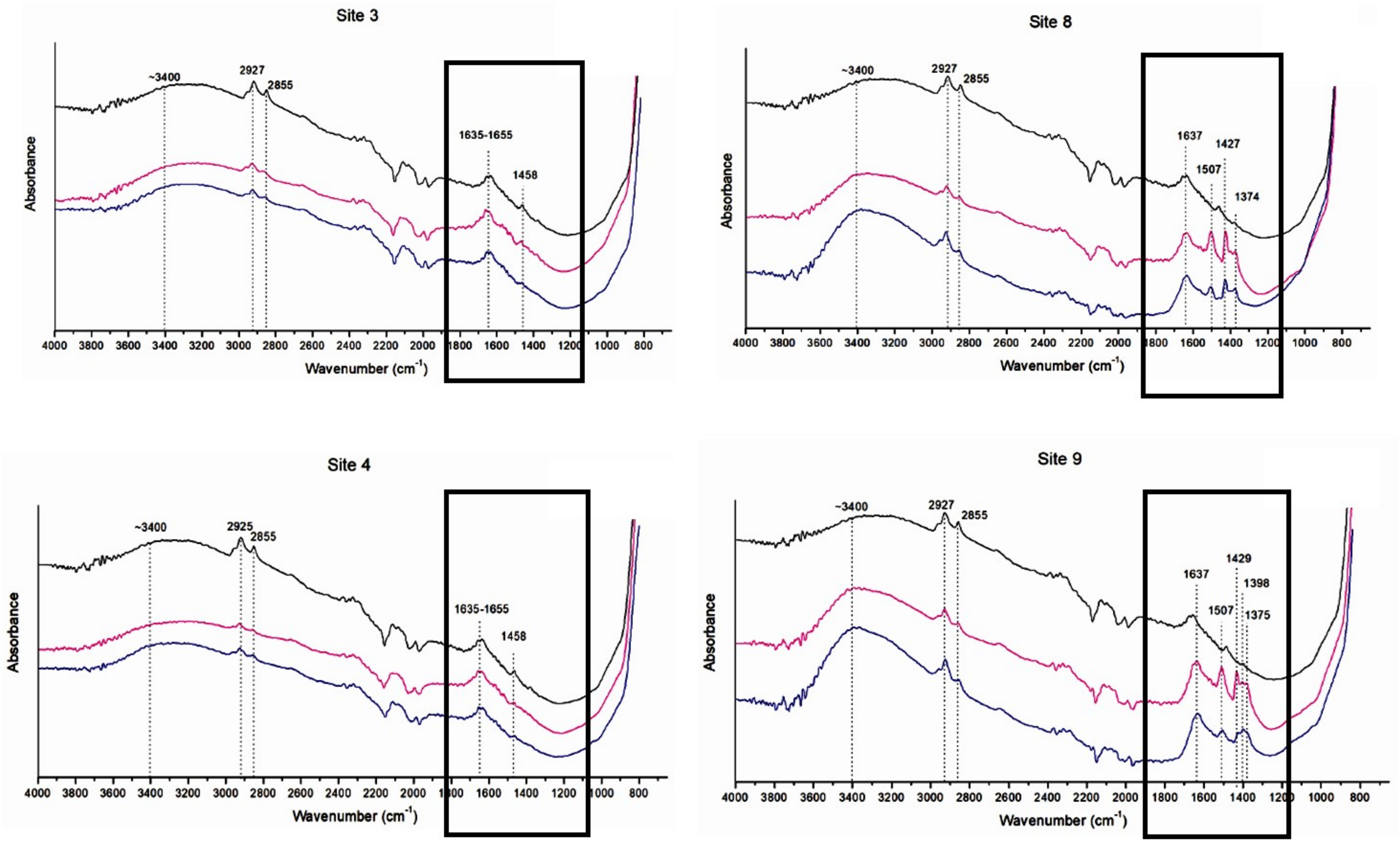

Figure S4: Representative ATR-FTIR spectra of the SW3, SW4, SW8, and SW9 in summer experiments. Black spectra: n-TiO_2, Red spectra: n-TiO_2/SWs, Blue spectra: n-TiO_2/SWs rinsed with pure water.

Table S3: Presence or absence of new ATR-FTIR bands on n-TiO_2 after exposure to SWs (n-TiO_2/SWs).

Site	Season	IR-wide1375-1398	IR-sharp-1430	IR-1500
SW1	**Spring**	X	-	-
	Summer	-	-	-
SW2	**Spring**	X	-	X
	Summer	X	-	-
SW3	**Spring**	-	-	-
	Summer	-	-	-
SW4	**Spring**	-	-	-
	Summer	-	-	-
SW5	**Spring**	X	-	X
	Summer	X	X	X
SW6	**Spring**	X	-	X
	Summer	X	X	X
SW7	**Spring**	X	X	X
	Summer	NA	NA	NA
SW8	**Spring**	X	X	X
	Summer	X	X	X
SW9	**Spring**	X	X	X
	Summer	X	X	X
SW10	**Spring**	NA	NA	NA
	Summer	X	-	-

X: presence of assigned ATR-FTIR band.

NA: lost sample/ samples could not be retrieved after exposure.

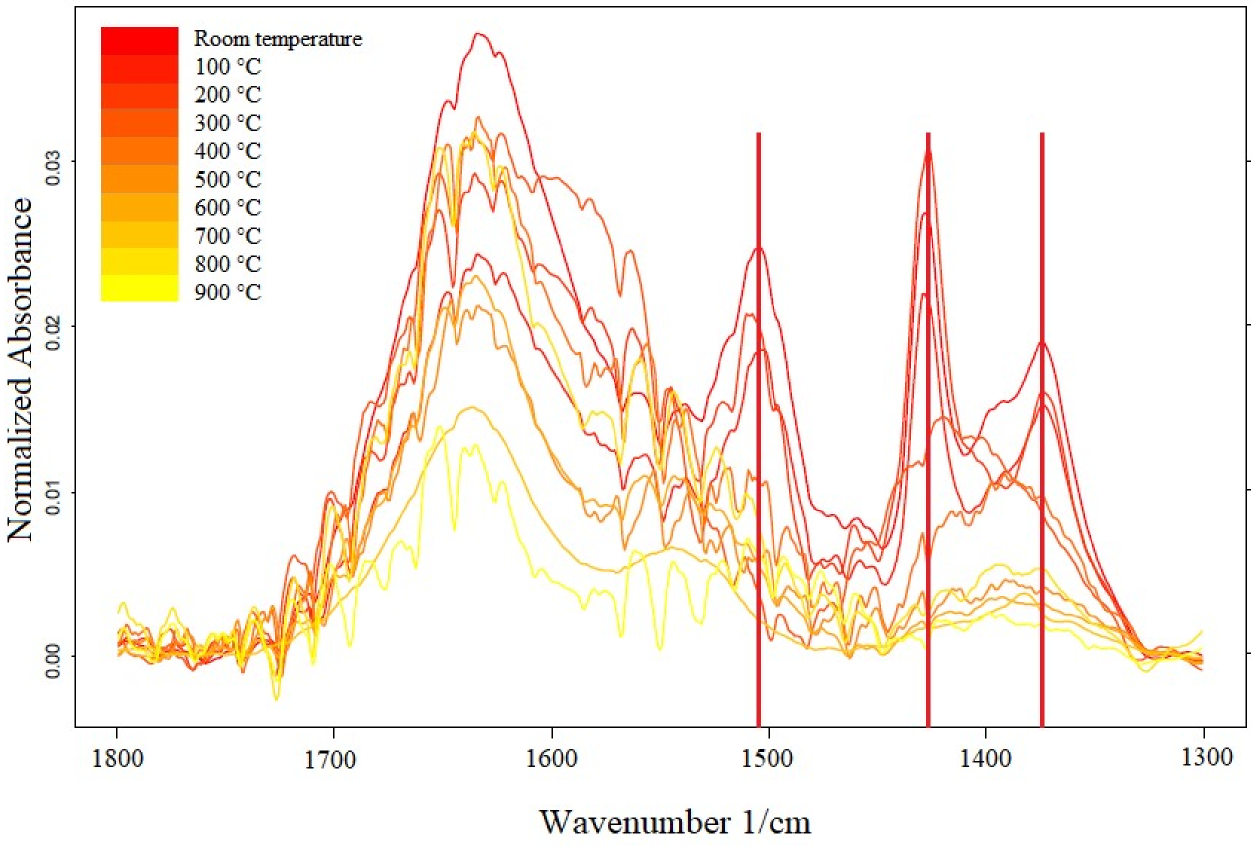

Figure S5: ATR-FTIR spectra of TiO_2/SW8 calcined stepwise from room temperature to 900 °C.

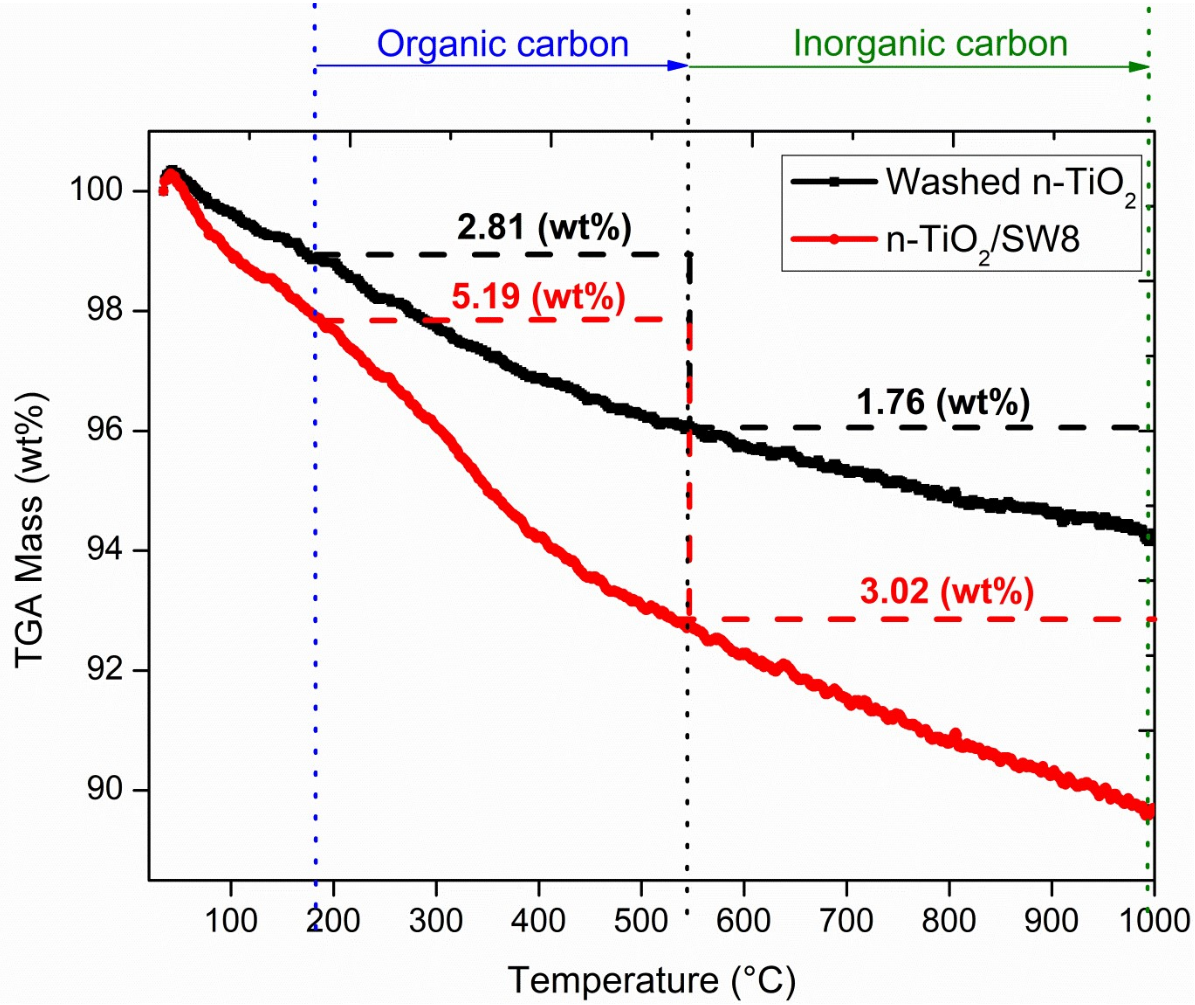

Figure S6: TGA curves for washed n-TiO_2 (control sample) and n-TiO_2/SW8.

Table S4: Atomic percentage of the elements present on-TiO_2/SWs samples in summer experiment obtained from XPS analysis.

Photoelectron line	Binding Energy (eV)	SW1 Atomic %	SW2 Atomic %	SW3 Atomic %	SW4 Atomic %	SW5 Atomic %	SW6 Atomic %	SW8 Atomic %	SW9 Atomic %	SW10 Atomic %	n-TiO2[1] Atomic %	Assigned bonds or functions
P $2p_{3/2}$	133.6	0.2	0.1	-	-	0.2	-	0.1	0.0	0.1	-	PO_4^{3-}
S $2p_{3/2}$	168.9	0.7	0.6	0.4	0.4	0.6	0.8	0.4	0.4	0.4	-	SO_4^{2-}
C 1s	285.0	24.4	26.2	26.5	26.9	26.5	27.7	28.5	27.0	26.9	19.5	C-C and C-H
C 1s	286.7	2.5	2.4	2.4	2.3	2.6	2.4	2.0	2.6	2.5	1.7	C-O
C 1s	288.6	1.4	1.7	1.8	1.4	1.9	1.9	1.5	1.7	1.8	1.8	O-C=O
C 1s	289.7	0.4	0.5	0.2	0.1	0.9	0.9	1.0	1.0	0.4	-	CO_3^{2-}
C total		*26.9*	*30.8*	*30.8*	*33.2*	*33.9*	*32.7*	*33.2*	*32.3*	*31.8*	*23.0*	
Ca $2p_{3/2}$	347.6	0.4	0.8	0.2	0.2	1.5	1.5	1.4	1.5	0.6	-	Ca^{2+}
Mg 1s	1304.4	0.1	0.2	0.1	-	0.3	0.7	0.2	0.3	0.2	-	Mg^{2+}
Ti $2p_{3/2}$	458.9	19.7	18.9	20.1	19.8	18.1	17.7	18.0	18.4	19.1	23.8	Titanium in TiO_2
O 1s	530.1	39.4	38.0	40.5	39.7	35.5	34.6	36.6	36.8	38.1	47.4	Oxygen in TiO_2
O 1s	531.9	10.0	9.9	7.3	8.0	11.4	11.2	9.6	9.8	9.4	5.6	Oxygen in C-O, O-C=O, CO_3^{2-}, PO_4^{3-}, SO_4^{2-}

[1] washed n-TiO_2 (control sample) before exposure to surface waters.

The data for SW7 is not available because the samples could not be retrieved after exposure.

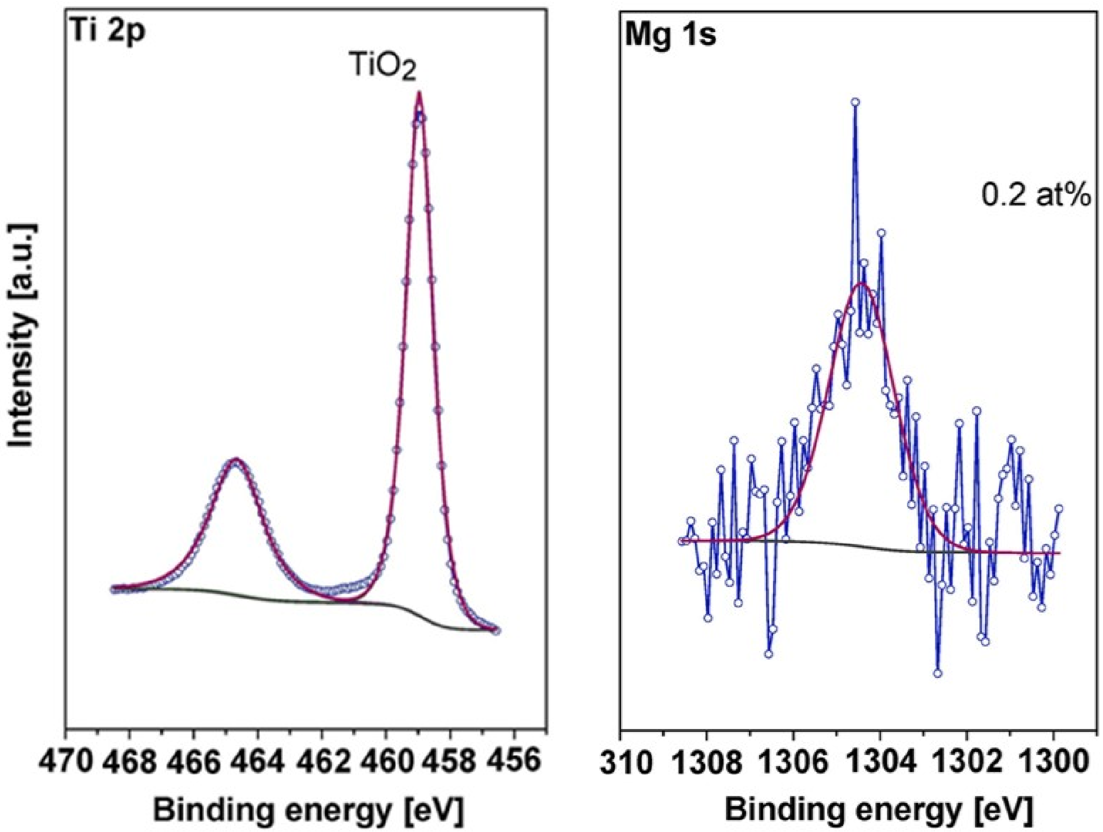

Figure S7: Illustrative Ti 2p and Mg 1s XPS peaks, baseline and fittings for n-TiO_2/SW8.

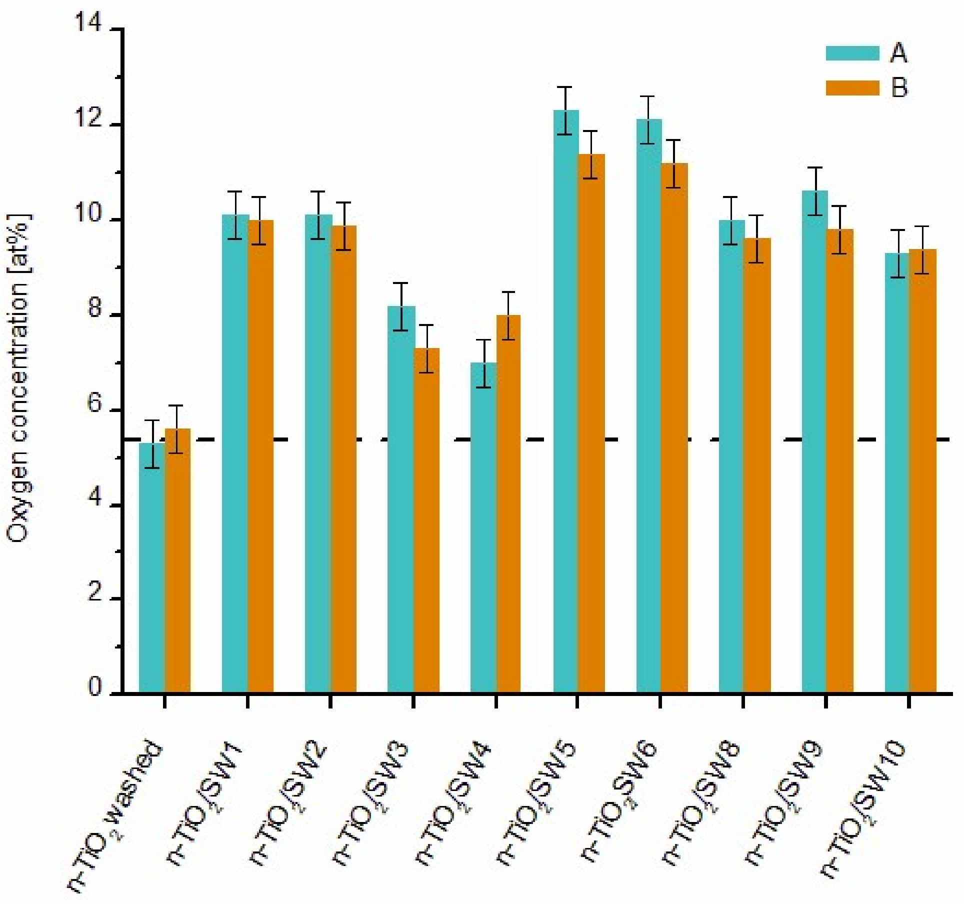

Figure S8: Comparison of oxygen content of n-TiO_2 exposed to SWs (n-TiO_2/SWs) in summer experiment. A) expected oxygen% in oxygen containing compounds calculated from C─O, O─C═O, and CO_3^{2-} (C 1s), SO_4^{2-} ($S2p_{3/2}$), and PO_4^{3-} ($P2p_{3/2}$), B) oxygen% measured

for O 1s at 531.9 eV. Sample SW7 could not be recovered. (The dashed line depicts the oxygen concentration (531.9 eV) in washed n-TiO_2).

C 1s and O 1s can confirm the presence of carbonate on the surface of nanoparticles. The C 1s XPS spectra of samples (n-TiO_2/SWs) exhibited three main peaks with binding energies around 285.0, 286.7, and 288.6 eV correspond to C─C (and C─H), C─O, and O─C═O bonds, respectively **(Figure 6 and Table S4)**. There was also a weak C 1s peak around 289.7 eV that can be assigned to CO_3^{2-}. This peak is not seen in the washed n-TiO_2 **(Table S4)**.

O 1s of the samples depict two peaks at 530.1 eV and 531.9 eV. The former (530.1 eV) is related to lattice oxygen in metal-oxides in accordance with the ratio of O/Ti = 2.0 in n-TiO_2 structure (**Table S4**). Interpretation of the later O 1s peak (531.9 eV) is not straightforward since binding energies of different oxygen containing compounds e.g. carbon-oxygen bonds in organic components as well as oxyanions fall in this range; hence, this peak may be partly due to surface oxygen ions in carbonate. If relative oxygen content measured for O 1s (531.9 eV) equals the relative oxygen content calculated from oxygen containing compounds i.e. C─O, O─C═O, CO_3^{2-} (C 1s), SO_4^{2-} (S $2p_{3/2}$), and PO_4^{2-} (P $2p_{3/2}$) the assignment of C 1s peak at 289.6 eV to CO_3^{2-} can be confirmed. The measured relative oxygen content oxygen% at 531.9 eV for each sample is depicted in **Table S4**. The calculation of oxygen% for oxygen containing compounds is performed using the expected stoichiometry and the resulting **equation 1**.

(1)

Oxygen% in oxygen containing compounds (calculated from C 1s, P $2p_{3/2}$, and S$2p_{3/2}$) =

1* carbon% in (C-O) + 2* carbon% in (O─C═O) + 3* carbon% in (CO_3^{2-}) + 4* phosphorus% in (PO_4^{3-}) + 4* sulfur% in (SO_4^{2-})

(Position of XPS peaks: C-O: 286.7 eV, O─C═O: 288.6 eV, CO_3^{2-}: 289.7 eV, PO_4^{3-}: 133.6 eV, and SO_4^{2-}: 168.9 eV)

Figure S8 shows the comparison of measured oxygen% at 531.9 eV and the expected oxygen% obtained from oxygen containing compounds (equation 1) for TiO_2/SWs in summer experiment. The oxygen concentration at 531.9 eV fits well with the correspondent species in equation 1; hence, the various peak attributions are supported.

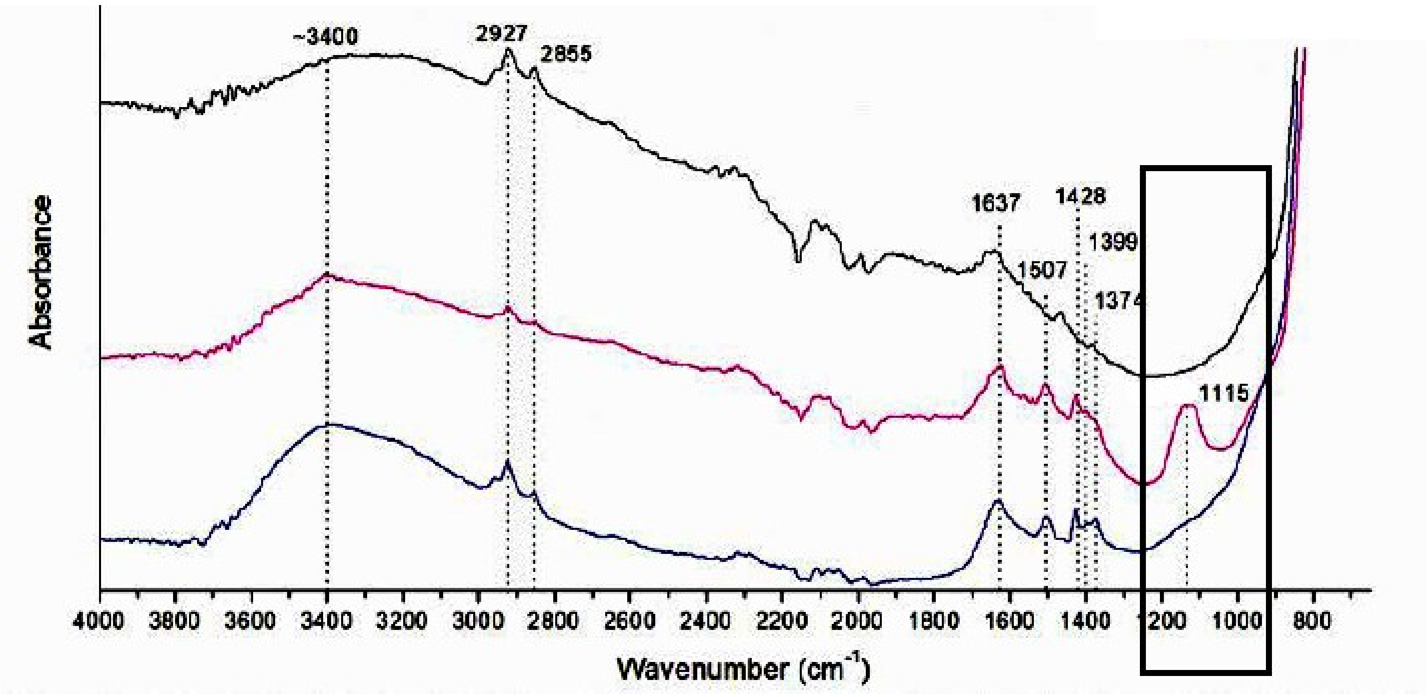

Figure S9: Representative ATR-FTIR spectra of n-TiO_2 exposed to SW6 in summer experiment. Black spectra: n-TiO_2, Red spectra: n-TiO_2/SW6, Blue spectra: n-TiO_2/SW6 rinsed with pure water.

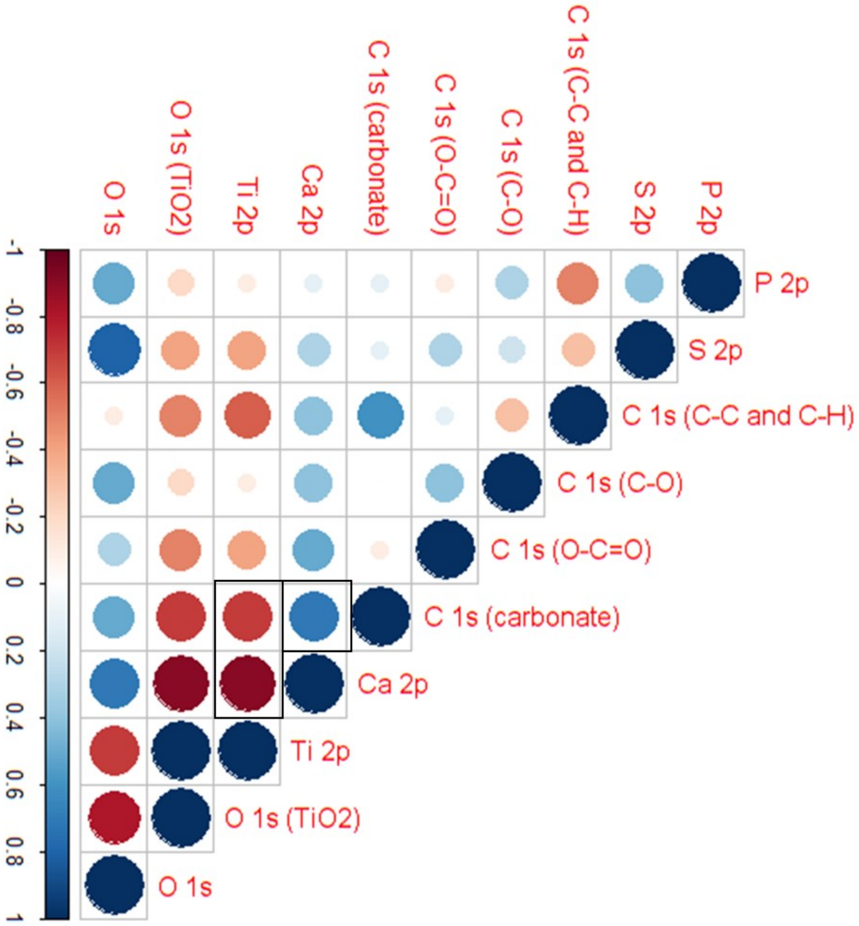

Figure S10: Correlation matrix among the elements and functional groups (analyzed by XPS) on the surface of n-TiO_2 exposed to surface waters (SW1-SW10). The color scale on the left denotes the correlation coefficient. The important correlations are framed.

Table S5: Zeta potential of n-TiO_2 /SWs samples in the summer experiment.[1]

TiO_2 nanoparticles exposed to surface waters	pH[2]	Zeta Potential (mV)
n-TiO_2 /SW1	6.8	-13.25 ± 1.02
n-TiO_2 /SW2	7.1	-27.18 ± 0.85
n-TiO_2 /SW3	6.5	-8.11 ± 0.57
n-TiO_2 /SW4	6.8	-11.53 ± 3.43
n-TiO_2 /SW5	8.2	-45.86 ± 1.73
n-TiO_2 /SW6	8.2	-23.11 ± 0.92
n-TiO_2 /SW8	8.0	-26.18 ± 1.02
n-TiO_2 /SW9	8.0	-25.15 ± 0.58
n-TiO_2 /SW10	7.2	-21.07 ± 1.67
Washed n-TiO_2	6.5	-6.04 ± 1.25

[1] The zeta potential was measured at 24 ºC with ionic strength of 0.01 M.

[2] The pHs of the samples are around the neutral pH, so they can be comparable.

Spearman's correlation matrix:

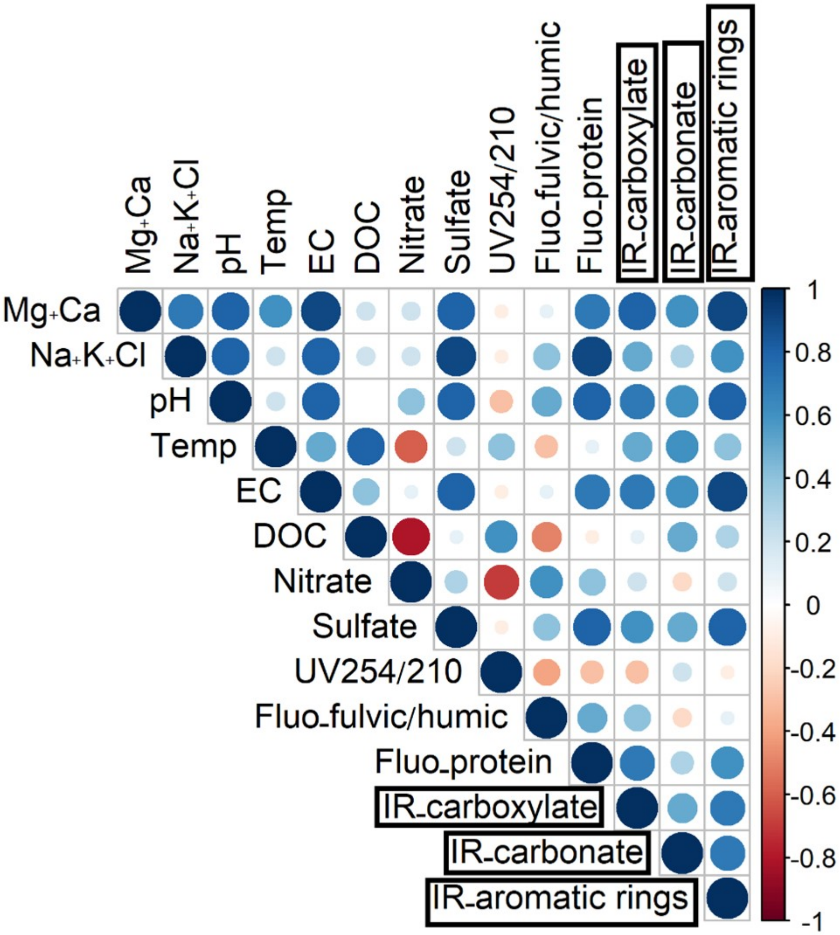

Figure S11: Correlation matrix among physicochemical parameters of SWs and functional groups appeared on nanoparticles after exposure to SWs. The color scale on the left shows (r) correlation coefficient. EC: electrical conductivity; DOC: dissolved organic carbon; Flu.

Protein: presence of a protein's fluorescence peak at Ex/Em ~ 270/330; Flu. fulvic/humic acid: the ratio of fluorescence intensity at Ex/Em ~ 340/430 to Ex/Em ~ 250/430. The functional groups assigned for IR bands are framed.

Figure S11 depicts the Spearman's correlation matrix among surface water parameters as well as sorbed groups (assigned from ATR-FTIR bands in Table 1) onto nanoparticles. The general correlations among surface water parameters are observed here as well. For instance, high correlation coefficients (r) are seen among EC and ionic contents such as Ca^{2+}-Mg^{2+} (r = 0.8), pH and EC are positively correlated (r = 0.8),[2] there is a negative correlation between DOC and NO_3^- concentrations (r = -0.8), etc.[3] The correlation matrix can be also used to recognize the degree of correlation between input (surface water parameters) and output (sorbed groups) variables.[4] Based on Spearman correlation, most of the variables are highly correlated (r ≥ 0.6) with sorbed groups. For instance, sorbed carboxylate groups are highly correlated to pH, and Ca^{2+}-Mg^{2+} (r = 0.7) of SWs which can depict the high influence of these parameters for the sorption of carboxylate groups. The highest correlation of carbonate is seen with EC and Ca^{2+}-Mg^{2+} (r = 0.6) in accordance with the presence of calcium carbonate on n-TiO_2/SWs.

Principal Component Analysis (PCA):

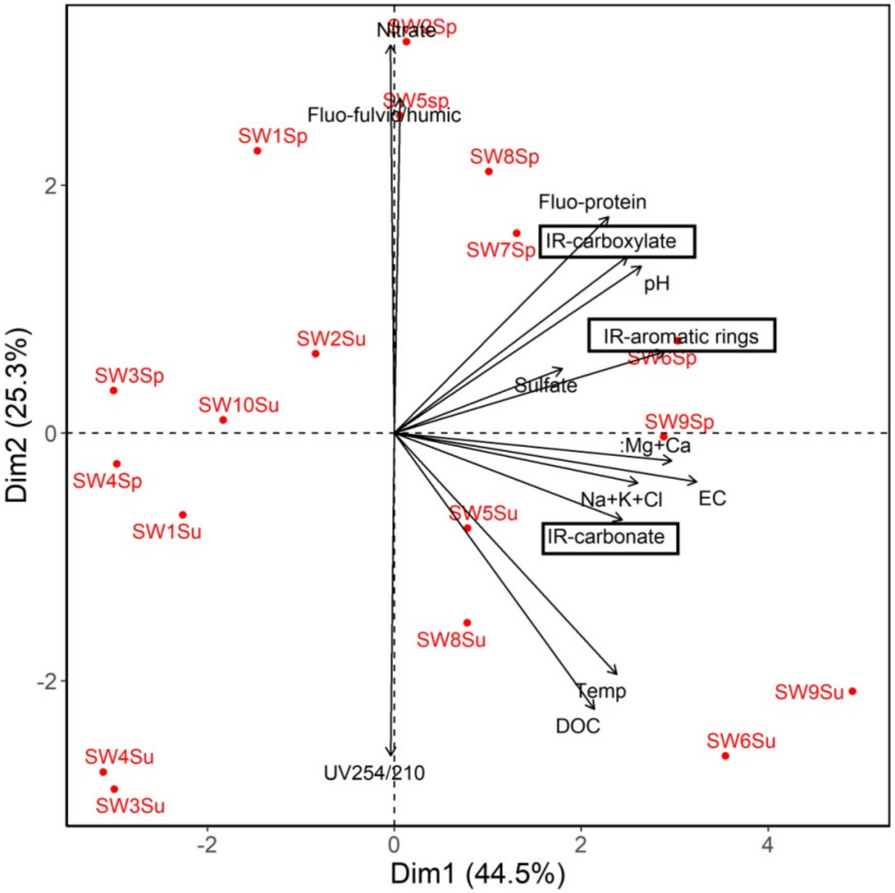

Figure S12: PCA biplot of individuals and variables. The biplot shows the PCA scores of variables (SWs physicochemical parameters) as vectors in black, and individuals i.e., samples in red. Spring and summer experiments are depicted by sp. and su., respectively. EC: electrical

conductivity; DOC: dissolved organic carbon; Flu. Protein: presence of a protein's fluorescence peak at Ex/Em ~ 270/330; Flu. fulvic/humic acid: the ratio of fluorescence intensity at Ex/Em ~ 340/430 to Ex/Em ~ 250/430. The functional groups assigned for IR bands are framed.

The principal component analysis (PCA) biplot of individuals and variables is represented in **Figure S12**. PCA axes, PC1 and PC2 together represented 65.4 % of variation of all the data set. Among samples, n-TiO_2 in SW5, SW6, SW7, SW8, and SW9 in both spring and summer experiments have high contribution on variable scores (vectors) since they are on the same side (right) as the given variables. Besides, the more parallel a variable vector to a principal component axis, the more it contributes to that component;[5] hence, Dim2 (or PC2, Eigenvalue= 25.3 %) mostly represents aromaticity factors of surface waters (UV 254/210 and Fluorescence fulvic/humic acid), and nitrate ions. On the other hand, Dim1 (or PC1, Eigenvalue = 44.5 %) represents sorbed/precipitated groups onto nanoparticles (carboxylate, carbonate, aromatic ring), and the correlated surface water parameters; which among them, EC has the highest contribution (longest vector). PCA can also show the relationships between dependent variables,[6] e.g. a high positive correlation is seen between "carboxylate" and pH, "aromatic ring" and Sulfate, and "carbonate" and Na^+-K^+-Cl^- ions. Similar to Spearman correlation, PCA depicts a low influence of NO_3^-, PO_4^{3-}, UV 254/210, and Flu. fulvic/humic acid (PC2) on composition of functional groups of natural coating (PC1). Also, a high negative correlation (opposite angles) between fluorescence fulvic/humic acid and UV 254/210 can be explained as a negative correlation based on the aromaticity of the system i.e. aromaticity of surface waters is proportional to UV 254/210,[7] and is inversely proportional to fluorescence fulvic/humic acid.[8]

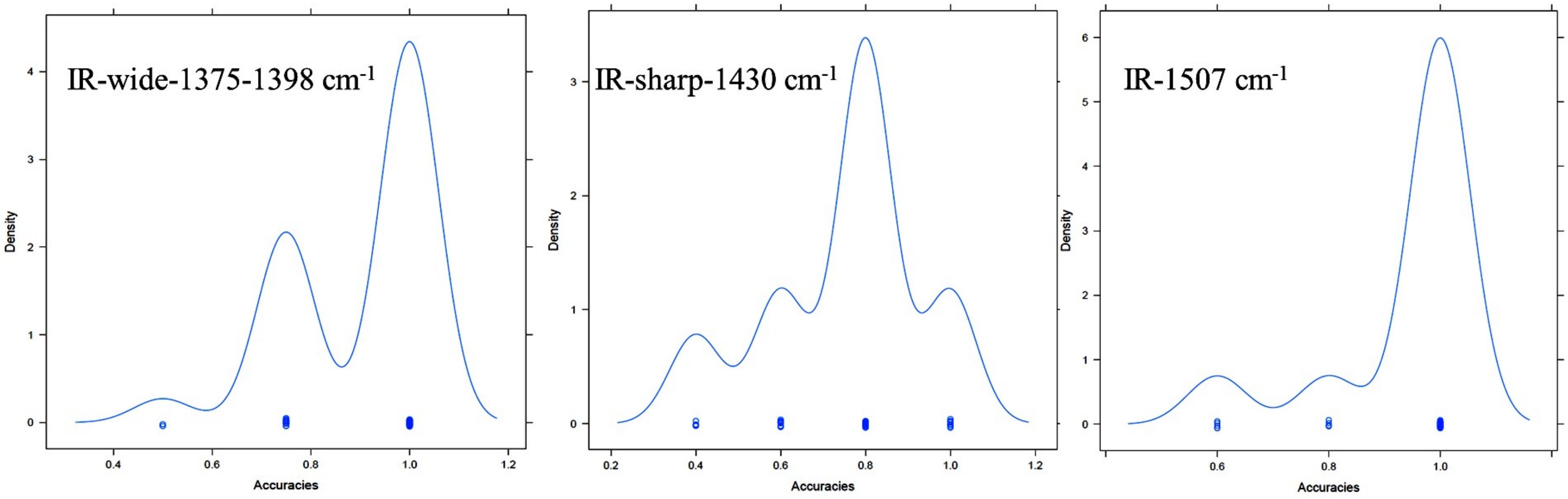

Figure S13: Density plots of the accuracy of the RF-models (50 models from 50 different initial data partitioning) for ATR-FTIR bands of n-TiO_2/SWs.

References:

(1) Tayyebi Sabet Khomami, N. T.; Philippe, A.; Quba, A. A. A.; Lechtenfeld, O.; Guigner, J.-M.; Heissler, S.; Schaumann, G. E. Validation of a Field Deployable Reactor for in-Situ Formation of NOM-Engineered Nanoparticles Corona. *Environmental Science: Nano* **2020**.

(2) Ratnayake, A.; Dushyantha, N.; De Silva, N.; Somasiri, H.; Jayasekara, N.; Weththasinghe, S.; Samaradivakara, G.; Vijitha, A.; Ratnayake, N. Sediment and Physicochemical Characteristics in Madu-Ganga Estuary, Southwest Sri Lanka. *J Geol Soc Sri Lanka* **2017**, *18*, 43–52.

(3) Konohira, E.; Yoshioka, T. Dissolved Organic Carbon and Nitrate Concentrations in Streams: A Useful Index Indicating Carbon and Nitrogen Availability in Catchments. In *Forest Ecosystems and Environments*; Springer, 2005; pp. 125–131.

(4) Jayaweera, C.; Aziz, N. Reliability of Principal Component Analysis and Pearson Correlation Coefficient, for Application in Artificial Neural Network Model Development, for Water Treatment Plants. In *IOP Conference Series: Materials Science and Engineering*; 2018; Vol. 458, p. 012076.

(5) Statistics and Geospatial data analysis https://www.geo.fu-berlin.de/en/v/soga/Geodata-analysis/Principal-Component-Analysis/principal-components-basics/Interpretation-and-visualization/index.html (accessed 2, 2021).

(6) Syms, C. Principal Components Analysis. In; Elsevier, 2008.

(7) Her, N.; Amy, G.; Sohn, J.; Gunten, U. UV Absorbance Ratio Index with Size Exclusion Chromatography (URI-SEC) as an NOM Property Indicator. *Journal of Water Supply: Research and Technology—AQUA* **2008**, *57*, 35–44.

(8) Sierra, M.; Giovanela, M.; Parlanti, E.; Soriano-Sierra, E. Fluorescence Fingerprint of Fulvic and Humic Acids from Varied Origins as Viewed by Single-Scan and Excitation/emission Matrix Techniques. *Chemosphere* **2005**, *58*, 715–733.

#R codes#

```
# libraries
# data cleaning/ data transformation
library(data.table)
library(dplyr)
library(tidyverse)

# cluster analysis
library(cluster)
library(factoextra)
library(optpart)

# correlation analysis
library(corrplot)
library(corrr)
library(corr2D)

#PCA
library(devtools)
library(factoextra)
library(chemometrics)

# random forest
library(rsample)
library(ranger)
library(randomForest)
library(vip)
library(caret)

# plotting
library(ggplot2)
library(cowplot)
library(grid)
library(ggcorrplot)


###################
### Import data ###
###################
Data <- read.csv(file.choose(), dec = ",", sep = "\t")

# Convert id as rownames
rownames(Data) <- as.character(Data[[1]])
Data[[1]] <- NULL
```

```
# Formating variables
Data[17:length(Data)] <- lapply(Data[17:length(Data)], FUN = as.factor)

########################
### Data Exploration ###
########################

# Scatter plots for exploring high correlations
plot(as.numeric(Data$Ca)~Data$Mg, xlab = expression(paste("Ca"^"+2")),
     ylab = expression(paste("Mg"^"+2")))
plot(as.numeric(Data$Na)~Data$K, xlab = expression(paste("Na"^"+")),
     ylab = expression(paste("K"^"+")))
plot(as.numeric(Data$Na)~Data$Cl, xlab = expression(paste("Na"^"+")),
     ylab = expression(paste("Cl"^"-")))

### Remove highly correlated variable:
### Combine Ca and Mg
### Combine K, Na, and Cl
### Season and PO4 removed (never important in preliminary RF-models)
### Remove F (low concentrations)

Data_red <- Data[-c(1,6,7,10:14)]
Data_red <- cbind(Mg_Ca = Data$Ca + Data$Mg,
                  Na_Cl_K =  Data$Cl + Data$Na + Data$K,
                  Data_red)
colnames(Data_red)[c(1:2,9:14)] <- c("Mg+Ca","Na+K+Cl","UV254/210","Fluo-fulvic/humic","Fluo-protein","IR-carboxylate","IR-carbonate","IR-aromatic rings")

# Plot parameter density distributions
tiff("Density distributions.tiff", width = 7, height = 5, res = 400, units = "in")
Data_red %>% keep(is.numeric) %>% gather() %>%
  ggplot(aes(value)) +
  facet_wrap(~ key, scales = "free") +
  geom_histogram()+
  theme_classic()
dev.off()

####################
### Correlation  ###
####################
# Create a correlation matrix
tiff("Correlation map.tiff", width = 7, height = 7, res = 400, units = "in")
Data_red %>% apply(., MARGIN = 2, FUN = as.numeric) %>% cor(., method ="spearman")
%>%
  round(.,0.5) %>% corrplot(., method="circle", type="upper", tl.col = "black")
dev.off()
```

```
############
### PCA  ###
############
### Check requirements
## Multivariate normality
# Plot a QQ-plot
chisqplot.multi <- function(m, main="QQ plot", ylab=expression(paste(chi^2, " Quantile"))){
  # n x p numeric matrix
  x <- as.matrix(m)
  # centroid
  center <- colMeans(x)
  n <- ncol(x)
  cov <- cov(x)
  # distances
  d <- mahalanobis(x,center,cov)
  s <- sort(d, index=TRUE)
  q <- (0.5:length(d))/length(d)
  par(las=1, cex=1.2)
  plot(s$x, qchisq(q,df=n), main=main, xlab="Ordered Mahalanobis D2", ylab= ylab)
  abline(a=0,b=1)
}
chisqplot.multi(Data_red[1:10])

# Normal after reducing the variables
# Look for outliers
# Plot score distances and othogonal distances
par(mfrow=c(1,2),cex=2)
pcaDiagplot(Data_red[1:10],
            princomp(Data_red[1:10],cor = TRUE),
            a =2, quantile=0.975
            )

# Perform PCA
res.pca <- Data_red %>% apply(., MARGIN = 2, FUN = as.numeric) %>% prcomp(., scale =
TRUE)
# Plot biplot
dimnames(res.pca$x)[[1]] <- rownames(Data_red)
tiff("biplot.tiff", width = 7, height = 7, res = 400, units = "in")
fviz_pca_biplot(res.pca, repel = TRUE,
                col.var = "black",
                col.ind = "red"
                )+
  theme_classic()+
  theme(panel.border = element_rect(colour = "black", fill=NA, size=1),
        axis.title.y  = element_text(size=16),
        axis.title.x  = element_text(size=16),
        title =  element_blank(),
        text=element_text(size=16)
        )
dev.off()


########################
### CLUSTER ANALYSIS ###
########################
# Simple hierarchical clustering

#Scale data
#Cluster analysis is done on all measured surface water parameters
#Remove IR bands
tiff("hclust.tiff", width = 10, height = 5, res = 400, units = "in")
Data[-c(1,18:20)] %>% dist(., method = "euclidean") %>% hclust(., method = "ward.D2") %>%
  plot(., cex = 1, main = "", xlab = "", sub = "")
dev.off()

############################
#####  RANDOM FOREST  ######
############################
# Fix names
colnames(Data_red) <- make.names(colnames(Data_red))
## Random forest analysis function
RF <- function(X,
               seed = 123,
               mtry_step = 1,
               label = NULL, Y) {
  # create training and test data set
  set.seed(seed)
  split <- initial_split(X, prop = .7, strata = Y)
  X_train <- training(split)
  X_test  <- testing(split)
  # number of features
  n_features <- length(setdiff(names(X_train), Y))
  # Grid for different parameters
  hyper_grid <-
    expand.grid(
      mtry        = seq(2, (length(X) - 1), by = mtry_step),
      node_size   = seq(1, 5, by = 1),
      sample_size = c(0.63, 0.7, 0.8),
      rmse = NA,
      OOB_error = NA
    )
  # Excute RF with grid search
  for (j in 1:nrow(hyper_grid)) {
    # train model
    model <- ranger(
      formula         = as.formula(paste0(Y,"~.")),
      data            = X_train,
      seed            = 123,
      verbose         = FALSE,
      num.trees       = n_features * 10,
      mtry            = hyper_grid$mtry[j],
      min.node.size   = hyper_grid$node_size[j],
      sample.fraction = hyper_grid$sample_size[j]
    )
    # add OOB error to grid
    hyper_grid$OOB_error[j] <- model$prediction.error
    hyper_grid$rmse[j] <- sqrt(model$prediction.error)
  }

  # top 50 models accoridng to OOB error
  Param_Optimization <-
    hyper_grid[order(hyper_grid$OOB_error), ] %>% head(., 50)
  # use best tuning parameters
  best_set <- hyper_grid[order(hyper_grid$OOB_error),][1,]
  # re-run model with permutation-based variable importance
  m_ranger_permutation <- ranger(
    formula         = as.formula(paste0(Y,"~.")),
    data            = X_train,
    num.trees       = n_features * 10,
    mtry            = best_set$mtry,
    min.node.size   = best_set$node_size,
    sample.fraction = best_set$sample_size,
    importance      = 'permutation',
    verbose         = FALSE,
    seed            = 123
  )
  # most important variables according to permutation method
  names(m_ranger_permutation$variable.importance) <- sub("_", ".",
                                                     sub(
                                                       "Mass_",
                                                       "",
                                                       names(m_ranger_permutation$variable.importance)
                                                     ))
  # access vip data
  var_imp <- vip::vip(m_ranger_permutation, num_features = 25, geom = NULL)
  RF_masses <- var_imp$data
  #### Predict for the test data ####
  pred_class <- predict(m_ranger_permutation,
                        X_test[, -which(names(X_test) == Y)]
                        )
  # Assess performance on test data
```

```
  Confusion_matrix <-
    caret::confusionMatrix(factor(pred_class$predictions),
                           factor(X_test[[which(colnames(X_test) == Y)]])
    )

  O <- list(
    Parameters_optimization = Param_Optimization,
    Important_masses = RF_masses,
    Confusion_matrix = Confusion_matrix
  )
  return(O)
}

# Run RF on 50 partitions for each IR-signal
Importances_all <- list()
k <- 1
for (j in 12:14){
  # Choose Data
  Data_RF <- Data_red[c(1:11,j)]
  Y <- names(Data_red)[j]
  # Random forest analysis for different data partitioning
  Results_RF <- list()
  Accuracies <- c()
  Importances <- tibble(Variables = sort(colnames(Data_RF)[1:(length(Data_RF)-1)]))
  for (i in 1:50){
    Results_RF[[i]] <- RF(Data_RF,
                          mtry_step = 1,
                          Y = Y,
                          seed = i)
    Accuracies <- c(Accuracies,
                    Results_RF[[i]]$Confusion_matrix$overall[[1]])
    Importances <- cbind(Importances,

Results_RF[[i]][["Important_masses"]][order(Results_RF[[i]][["Important_masses"]]$Variable),
2]
    )
  }

  # Plot accuracies for the 50 partitions
  tiff(paste0("Accuracies-",Y,".tiff"),width = 5, height = 5, units = "in", res = 400)
  print(lattice::densityplot(Accuracies))
  dev.off()

  #Save Importances for the 50 partitions
  rownames(Importances) <- Importances$Variables
  Importances$Variables <- NULL
  Importances_all[[k]] <- Importances %>% t(.) %>% as.data.frame(.) %>% cbind(.,Y)
  k <- k+1
}
Importances_all <- rbind(Importances_all[[1]],
                        Importances_all[[2]],
                        Importances_all[[3]])

tiff(paste0("Importances.tiff"), width = 10, height = 7, res = 400, units = "in")
par(mar=c(8.1, 4.1, 2.1, 1.1),las=2,cex.lab = 1.25,cex.axis = 1.1)
boxplot(Importances_all[,-length(Importances_all)], boxfill = NA, border = NA,
        ylab = "Importance",
        names = c("DOC","EC","Flu. fulvic/humic","Flu.
Protein","Ca+Mg","Na+K+Cl","Nitrate","pH","Sulfate","T°C","UV 254/210")
)
boxplot(Importances_all[Importances_all$Y=="IR.carboxylate", -length(Importances_all)],
        xaxt = "n", yaxt = "n",
        add = TRUE,
        boxfill="red3",
        boxwex=0.19,
        at = 1:ncol(Importances_all[,-length(Importances_all)]) - 0.23
)
boxplot(Importances_all[Importances_all$Y=="IR.carbonate", -length(Importances_all)],
        xaxt = "n", yaxt = "n",
        add = TRUE,
        boxfill="orangered",
        boxwex=0.19,
        at = 1:ncol(Importances_all[,-length(Importances_all)])
)
boxplot(Importances_all[Importances_all$Y=="IR.aromatic.rings", -length(Importances_all)],
        xaxt = "n", yaxt = "n",
        add = TRUE,
        boxfill="yellow",
        boxwex=0.19,
        at = 1:ncol(Importances_all[,-length(Importances_all)]) + 0.23
)
legend("topright",
       legend=c("IR-carboxylate", "IR-carbonate", "IR-aromatic"),
       fill=c("red3", "orangered","yellow"),
       bty = "n", cex = 1.1)

dev.off()

#########
##2D-IR##
#########

#Reading
read_ftir <- function(path = ".", dec = ".", encoding = getOption("encoding")) {
  files <- list.files(path = path, pattern = "\\.asp$")
  ret <- data.frame()
  for (cfile in files) {
    con <- file(file.path(path, cfile), encoding = encoding)
    lines <- as.numeric(readLines(con))
    close(con)
    sig <- lines[-c(1:6)]
    dat <- rbind(sig)
    colnames(dat) <- as.character(seq(lines[2], lines[3], length.out = lines[1]))
    rownames(dat) <- sub("\\.asp$", "", cfile)
    ret <- rbind(ret, dat)
  }
  return(ret)
}
#SG-smoothing
Smoothing <- function(X){
  Spectra_smoothed <- X
  for (i in 1:length(Spectra_smoothed[[1]])){
    Spectra_smoothed[i,] <- signal::sgolayfilt(as.data.frame(t(X[i,]))[[1]], 3, 21)
  }
  return(Spectra_smoothed)
}
#Normalizing
Normalizing <- function(X, method="TiO2"){
  if (method=="TiO2"){
    Y <- X
    for (i in 1:length(X[[1]])){
      Y[i,] <- apply(Y[i,],MARGIN = 1, FUN = function(Z){Z/Y[i,length(Y)]})
    }
    return(Y)
  }
  if (method=="scale"){
    Y <- X %>% as.matrix(.) %>% t(.) %>% scale(.,center=T,scale=T) %>% t(.) %>%
as.data.frame(.)
    return(Y)
  }
  if (is.null(method)==T){}
}
#Selecting a window
Selecting <- function(X, START = 0, END = length(X), SAMPLE_EXCL = NULL){
  if (is.null(SAMPLE_EXCL) == T){
    Y <- X[which(as.numeric(colnames(X))>START &
                   as.numeric(colnames(X))<END)]
  } else {
    Y <- X[-SAMPLE_EXCL,
           which(as.numeric(colnames(X))>START &
                   as.numeric(colnames(X))<END)]
  }
  return(Y)
}
#Removing trends
##!!!has to be optimize for each range!!!##
##For 1400-1800: df:5; for 2700-3100: df=3;iterations=5;threshold=0.01##
Remove_baseline <- function(X, df = 3, iterations = 5){
  Y <- X
  for (i in 1:length(X[[1]])){
    w <- rep(1,times = length(X))
    for (j in 1:iterations){
      Trend <- X[i,] %>% smooth.spline(x=as.numeric(colnames(X)),
                                       y=.,
                                       w=w,
                                       df=df)
      w[which(X[i,] - predict(Trend, as.numeric(colnames(X)))$y > 0.01*mean(t(X[i,])))] <- 0
    }
    Y[i,] <- X[i,]-predict(Trend, as.numeric(colnames(X)))$y
  }
  return(Y)
}

#####################
###Import IR Data#####
#####################
#Define path and process data
path <- "file.choose"
path %>% read_ftir(.) %>% Smoothing(.) %>% Normalizing(.,method = "TiO2") ->
Spectra_processed
#%>% Remove_baseline(.,df=5,iterations=10)
path %>% read_ftir(.) %>% Smoothing(.) %>% Normalizing(.,
                                                       method = "TiO2"
) %>% Selecting(.,
                START = 1300,
                END = 1800
) %>% Remove_baseline(.,
                      df=3,
                      iterations=5
) -> Spectra_processed_1
path %>% read_ftir(.) %>% Smoothing(.) %>% Normalizing(.,
                                                       method = "TiO2"
) %>% Selecting(.,
                START = 2750,
                END = 3050
) %>% Remove_baseline(.,
                      df=3,
                      iterations=5
```

```
) -> Spectra_processed_2

Both_regions <- path %>% read_ftir(.) %>% Smoothing(.) %>% Normalizing(.,method =
"TiO2") %>% Selecting(.,1300,3050)
Both_regions[which(as.numeric(colnames(Both_regions)) < 1800)] <- Spectra_processed_1
Both_regions[which(as.numeric(colnames(Both_regions)) > 2750)] <- Spectra_processed_2

#Plot all spectra
plot_spectra <- function(X, Start = 1, End = length(X[[1]])){
  X %>% apply(X = .,MARGIN = 1,FUN = max) %>% max(.) -> MAX
  X %>% apply(X = .,MARGIN = 1,FUN = min) %>% min(.) -> MIN
  plot(as.data.frame(t(X[1,]))[[1]]~
         as.numeric(colnames(X)),
       type = "n",
       ylim = c(MIN,MAX),
       xlab = expression("Wavenumber in cm"^-1),
       ylab = "Normalized Absorbance")
  pal <- colorRampPalette(c("red", "yellow"))
  for (i in Start:End){
    lines(as.data.frame(t(X[i,]))[[1]]~
            as.numeric(colnames(X)),
          type = "l",
          col = pal(End)[i],
          lwd = 2, legend(),
  }
}
plot_spectra(Spectra_processed_2)



#Synchronous
plot_corr2d(Corr_Spectra,
            xlim = c(1300,1550),
            ylim = c(1300,1550)
)


#Asynchronous
plot(Corr_Spectra, Im(Corr_Spectra$FT),
     xlim = c(1300,1550),
     ylim = c(1300,1550)
)
#Asynchronous
plot(Corr_Spectra, Im(Corr_Spectra$FT),
     xlim = c(1300,1530),
     ylim = c(2750,30000)
)
```