Measuring and predicting the ζ-potential of anthropogenic TiO_2-nanoparticles in surface waters

Allan Philippe [a*], Narjes Tayyebi Sabet Khomami [a], Michel Gad [b], Fintan Hahn [a], Vanessa Trouillet [c], Oliver Lechtenfeld [b], Stefan Kunz [d], María José Gormaz Aravena [a], Vanessa Wollersen [e], Eliana Di Lodovico [a]

* Corresponding author

[a] iES Landau, Institute for Environmental Sciences, Group of Environmental and Soil Chemistry, University of Kaiserslautern-Landau (RPTU), 76829 Landau, Germany.

[b] Department of Environmental Analytical Chemistry, Helmholtz Centre for Environmental Research – UFZ, 04318 Leipzig, Germany.

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

[d] Department of Groundwater Area Information, Federal Institute for Geosciences and Natural Resources, 13593 Berlin, Germany.

[e] Institute of Nanotechnology (INT) and Karlsruhe Nano Micro Facility (KNMFi), Karlsruhe Institute of Technology (KIT), 76344 Eggenstein-Leopoldshafen, Germany.

Abstract

We propose a novel approach for determining and predicting the ζ-potential of nanoparticles in surface waters based on the water composition and environmental parameters. Applying the dialysis bag method, five different types of TiO_2-nanoparticles representing the most common TiO_2-particles in commercial products were exposed in situ to a set of representative surface waters. ζ-potentials of these environmentally coated particles ranged from -58 to 13 mV and were used together with water composition data to train models for predicting the ζ-potential from the water composition. With an average root mean square error of 3.6 mV for 50 generated models, the XGBoost models outperformed random forest and various linear

models. We explored these models using parameter importances and shap values. Furthermore, we characterized the surface coating of a selection of samples using XPS and TG-QMS. Using these techniques, we could confirm the presence of an organic coating and explore the connections between ζ-potential values and environmental coating. As expected from batch experiments studies, the concentration of divalent cations is the most important factor for predicting ζ-potential of environmentally coated nanoparticles in surface waters. We found that the quality of dissolved organic matter has a significant effect whereas pH and dissolved organic matter content were less important. This study demonstrates the potential of our in-situ exposure method combined with multivariate analysis to explore the fate of nanoparticles in aquatic environments.

Keywords

titanium dioxide, colloid, surface coating, random forest, in-situ, eco-corona

Introduction

Titanium dioxide (TiO_2) is present in the form of nanoparticles in sunscreens, construction materials, food, and paints, for instance, with the role of UV-filter, photocatalyst, and white pigment, depending on the product.[1] In paints and coating, its high refractive index and low-cost have made TiO_2 the most popular choice as a white pigment.[2] The release of particles from construction materials and coatings exposed to weathering is an issue which has recently raised awareness due to the various environmental impacts of these particles.[3] Furthermore, the presence of anthropogenic TiO_2 in runoff from urban infrastructures during rain events was observed on-site in several studies suggesting that its concentrations in the environment are increasing.[4–7] TiO_2-particles could also be released indoors as suggested by their presence in home-dust.[8] More direct release of TiO_2 has been observed in various bathing waters due to the wash-off of sunscreens from bathers.[1] Hence, anthropogenic TiO_2-particles have been detected in rivers[9–13] and lakes.[14,15] These considerations result in concerns about their

environmental impact. Indeed, despite its relatively low toxicity,[16,17] TiO_2 is non-biodegradable and accumulates in surface waters and sediments over time.[1,18] Therefore, it cannot be excluded that anthropogenic TiO_2 will have significant negative impacts on the environment in the near future.

Therefore, it is important to understand the fate of nanoparticles in surface waters for predicting in which areas they could potentially accumulate and be of concern. It is already well accepted that the most important flux is the transport in rivers.[18,19] In this medium, the nanoparticles are suspended and their settling, uptake, and attachment rate depend on their surface properties. For instance, it is established that aggregation rate depends on surface charge, surface coating composition, and physical structure.[20–22] The modelling of these processes using, for instance, the DLVO-theory requires the ζ-potentials of the aggregating particles as input to predict the attachment efficiency.[23]

The ζ-potential is defined as the electrostatic potential at the slipping plane below which the charged solutes cannot diffuse freely anymore and are moving together with the particle.[24] Its value depends, therefore, on the surface composition of the particle, the quantity and quality of the electrolytes present in the medium, the pH, and the temperature.[24] Since the surface composition is strongly affected by the synthesis method, the ζ-potential will be different for different types of TiO_2 as suggested by isoelectric points ranging from 5.3 to 7.7.[25] Furthermore, TiO_2-particles used in commercial products are most of the time coated to tune their chemical properties. For instance, the surface can be passivated using Al_2O_3 or SiO_2 to minimize photocatalytic activity in cosmetic products[26,27], or organic polymers can also be added to increase the hydrophobicity of the particles.[2,27,28] This coating also influences the ζ-potential of the particles[27] and may experience drastic changes under environmental conditions.[26]

Once exposed to surface water, the ζ-potential will change in a complex and dynamic way due to the presence of electrolytes and dissolved organic matter (DOM) forming a complex environmental coating on the surface.[21,29,30] Unfortunately, most related environmental studies have been conducted with standard materials under controlled simplified conditions,[21,29,30] and it is therefore unclear if these results can be extrapolated to real-life nanoparticles under realistic environmental conditions. Furthermore, there is currently a lack of models for the prediction of ζ-potential in surface water which accounts for the complexity of the interactions between the relevant parameters.

In this study, we tackled these challenges by exposing five TiO_2-nanoparticles used in commercial products to a representative set of surface water *in situ* using a dedicated method developed in our group.[31,32] This *in situ* exposure procedure enables us to obtain particles with a realistic environmental coating. The ζ-potential of these particles in the surface waters was then determined and combined with the data from the water composition to obtain a dataset used to train multivariate models to predict ζ-potential from the water parameters. Analyzing these models and combining them with published results obtained under controlled conditions enables us to gain insight into the physico-chemical processes dominating the surface charge of TiO_2-particles in surface waters and, thereafter, to support more accurate fate modeling.

Material and methods

Nanoparticles

Five TiO_2 nanoparticles were selected based on their potential for environmental release, rather than to represent a diverse range of surface chemical compositions. Aeroxide® n-TiO_2 P25 powder was purchased from Evonik (Essen, Germany). The widely used photocatalyst P25 is composed mainly of anatase (~75%) and rutile (~25%) with size around 30 nm.[33] Furthermore, these particles were used in our previous studies describing the environmental coating formed under field conditions.[31,32] This enabled us to have a benchmark to compare

the results of this study with previous ones. The characterization of these particles can be found elsewhere.[31–33] We used the manufactured nanomaterial NM103 from the Joint Research Center of the European Commission. The rutile particles have a median diameter of 97 nm with a median aspect ratio of 1.6 and an Al_2O_3-coating and a PDMS coating.[34] These particles are similar to the ones used in cosmetics, in particular sunscreens. A white pigment commonly used in paint formulations: R-TC90 produced by Tioxide® was purchased from Sehestedter Naturfarben Handel GmbH (Sehestedt, Germany). These rutile particles have a median length of 278 nm with a median aspect ratio of 1.36 and an Al-based coating (most probably AlO(OH)) and a hydrophobic organic coating.

We extracted TiO_2-nanoparticles from commercial sunscreen following a similar procedure as described elsewhere.[27] For that, 10 g of sunscreen (Alverde Naturkosmetik, DM, Germany) and 200 mL of n-hexane (>98.5%, 33321, Alfa Aesar, Karlsruhe, Germany) were stirred until a homogenous milky suspension was obtained. An aliphatic solvent was selected due to the highly hydrophobic nature of the sunscreen, which contains a substantial proportion of oil components. The suspension was transferred to 50 mL glass ultracentrifuge tubes and centrifuged at 5 000 rpm (~ 4 053 g) for 60 min (Universal 320, Hettisch, Germany). The supernatant was carefully removed with a glass pipette and the solid residue was redispersed in around 40 mL of hexane and centrifuged. This rinsing step was repeated one more time. The supernatant was then discarded, and the samples were left to dry overnight at ambient temperature. We added 40 mL of ultrapure water (UPW, Purelab Flex Elga-Veolia, 18.2 MΩ cm) and sonicated the suspension in a bath sonicator (USC1200TH, VWR, Leuven, Germany) with occasional stirring for around 5 min. The suspension was transferred to polypropylene centrifuge tubes and centrifuged at 10 000 rpm (~14 087 g) for 30 min (Megastar 4.0, VWR, rotor HIGHConic). The supernatant was discarded and the samples dried at 60°C overnight. The average yield of the extraction was 1.66 ± 0.034 g corresponding to around 17% of the total sunscreen mass. The characterization of these extracted particles is described below.

We also extracted particles from commercial paint. The paint (Latexweiss, Schöner Wohnen, Germany) was sampled from the original recipient after vigorous shaking. We weighted 10 g of paint in a glass beaker and added 1L of 20 mM EDTA solution (UPW and 1.08418.0100 Merck Titriplex III, Darmstadt, Germany). The EDTA served as a complexing agent to increase the solubility of $CaCO_3$; one of the paint's main components. The suspension was stirred magnetically until it was homogeneous (~15 min, 500 rpm), then 16 g of NaOH pellets (>98%, p.a. ISO, 6771.1, Roth, Karlsruhe, Germany) were added and the suspension was warmed up at ~95°C for 3 h to dissolve the amphiphilic polymers in the paint. It was then transferred to a separation funnel and left for decanting for ~1 h. The particles formed a white milky layer at the lower part of the funnel. This fraction was transferred to 50 mL PP centrifuge tubes and centrifuged (Universal 320, Hettisch, Germany) at 5 000 rpm for 5 min. The supernatant was carefully pipetted out and 50 mL of UPW was added. The suspension was vortexed and centrifuged at 10 000 rpm for 15 min (Megastar 4.0, VWR, rotor HIGHConic). The rinsing with UPW was repeated three times in total. The final centrifugate was oven-dried at 60 °C for 24h. The average yield of the extraction was 2.9 ± 0.47 g, corresponding to around 29% of the total paint mass. Further details about the extraction method and the characterization of these particles, which are very similar to R-TC90, can be found elsewhere.[35]

Screening for exposure sites

We listed the sites to be screened using the FORGES database[36] by defining areas of interest based on the Ca^{2+}, Mg^{2+}, carbonate, Na^{+}, pH, sulfate, and total organic carbon (TOC) values. The selection was made to maximize the range of parameters covered by the areas and to have some contrasting compositions, e.g., high Ca^{2+} but low Mg^{2+}. In each area, locations were selected using Google Maps to find disparate sites in terms of the landscape around the water bodies or watersheds for rivers. Additionally, sites were included where detailed knowledge

of environmental parameters and monitoring data were available. In total, we sampled 84 surface waters across France, Switzerland, and Germany between May 11th and June 30th of 2021. The description of the methods used to sample and characterize the water as well as the obtained dataset can be found elsewhere.[37] Fluorescence excitation/emission spectra were recorded on a PerkinElmer LS55 Fluorescence spectrometer.[31] Fluorescence indices were calculated using the Stardom R package after correcting for the inner filter effect,[38] based on Coble indices. Fluo_b, Fluo_t, Fluo_a, Fluo_m, and Fluo_c indicate tyrosine-like, tryptophan-like, fulvic-like, marine-like, and humic-like fluorophores, respectively. Moreover, Fluo_bix, Fluo_fi, and Fluo_hix denote freshness index, fluorescence index, and humification index, respectively.[39] The interpretation of these indices can be found in Tables S1 and S2.

Selection of exposure sites

Using the screening data, we determined the 20 most representative sampling sites from the screening sites (n = 84). We first delineated groups of sites exhibiting similar values in their environmental variables using cluster analysis. For that, we calculated the Euclidean distance matrix from the z-transformed data and applied hierarchical cluster analysis (Ward's method) to it while maximizing the average silhouette width[40] to obtain the optimal number of clusters. We defined the closest site to the centroid (squared distance) of each cluster as the best representative for the cluster. This resulted in 19 clusters and representatives. When the scores of the closest sites to the centroid were equal or very close, the geographically closest or the most accessible site was selected. We used multidimensional scaling (MDS) on the scaled dataset using the Manhattan method (vegan R package) to confirm the representativity of our selection. The representative points were well distributed and covered most of the projected space, which confirms that our selection method was efficient. In addition to these sites, we decided to add site 51 as its distance to the centroid was the highest among the dataset and MDS analysis showed that site 51 filled a gap in the projected space (Figure S1). The screening sites and the final selection are shown in Figure 1.

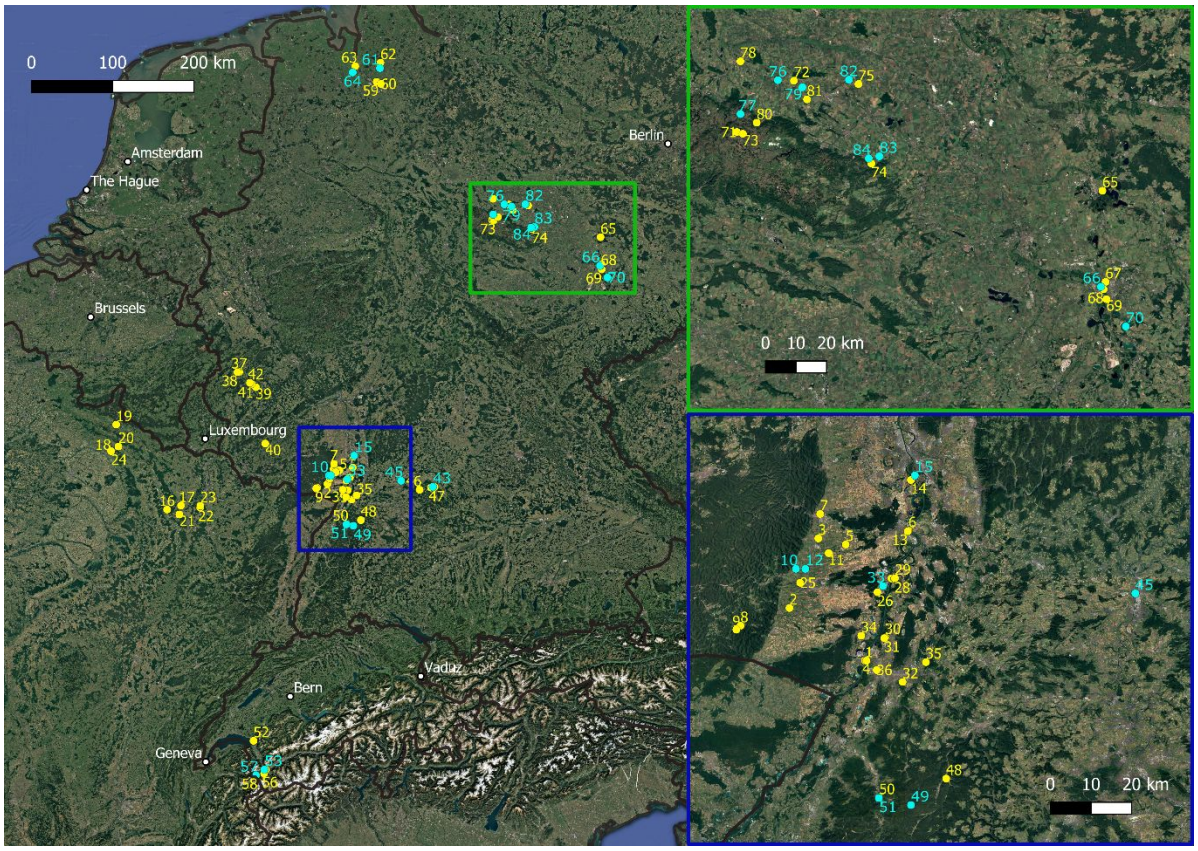


Figure 1: map of the sites screened in this study (yellow dots) and the sites selected for the exposure experiments (cyan dots). Each site was identified with a number. The map was generated using QGIS with Google satellite images.

Exposure experiments

Biotech cylindrical Cellulose ester (CE) Membranes (Float-A-Lyzer®, molecular weight cut off = 100 kDa, 10 mL) were purchased from Repligen (California, USA). The specifications of the dialysis bags, as well as the conditioning method, can be found elsewhere.[31] In brief, the purchased membranes were soaked in 10 % (v/v) ethanol-water and rinsed with distilled water (DW). The membranes were stored in DW until further usage.

The TiO_2 suspensions (2 g/L) were freshly prepared by sonicating the corresponding nanoparticule powder in UPW until a homogeneous suspension was obtained. The dialysis bags were filled with 10 mL of either distilled water (control) or a TiO_2 suspension (two replicates per TiO_2 type). Consequently, each dialysis bag was enclosed in a 3D-printed protective shell (Figure S2) which was fixed to an anchor and immersed in the surface water

about 1 m from the river/lake bank. A temperature sensor (Pendant MX, HOBO, Bourne, USA) was attached to the anchor for monitoring the temperature over the whole exposure duration. After 48h, the content of the dialysis bags was transferred into a 15 mL polypropylene centrifuge tube and centrifuged at 4 500 rpm (3 283 g) for 30 min. The supernatant was withdrawn carefully to the last drop using a plastic pipette. The remaining centrifugate was freeze-dried (Christ, Osterode, Germany) for two continuous days at -40 °C and 0.12 mbar. The freeze-dried powders were then stored in the dark at -20°C until further use.

The exposure experiments were carried out on each of the 20 selected exposure sites at three different seasons: April-Mai, July-August, and October-November 2023 except for site 55 for which the spring exposure was carried out in April 2024. For each exposure experiment, the surface water was sampled and analyzed following the same procedure as for the screening samples. It is important to note that a mislabeling incident resulted in the exclusion of half of the samples for P25.

Scanning electron microscopy (SEM)

The morphology of the TiO_2-particles was characterized by SEM (Zeiss Auriga60, Zeiss, Oberkochen, Germany). The already extracted particles were dispersed in deionised water by using an ultrasound bath for several minutes. For each sample a drop of suspension was transferred to a TEM grid (holey carbon film on Au-Grid) lying on filter paper to soak the excess water. By using the STEM grid holder, the samples were loaded into the SEM and imaged using 5 kV as high tension and the SESI (secondary electron/secondary ion) detector.

X-ray photoelectron spectroscopy (XPS)

The measurements were performed to confirm changes in surface chemistry after exposure. Due to cost and time limitations, we selected five sites in Germany with contrasting features from fall and summer for the XPS-measurements. Sites 10 and 12 were situated before and after a wastewater treatment plant on a stream sourcing from a large, forested area close to

Landau in der Pfalz and had contrasted DOM composition. Site 33 is a bathing lake close to the Rhine River in Germersheim. Site 47 was located on the stream Arzbach near Michelfeld in a highly calcareous region. Finally, site 49 was from a moor in the Black Forest.

XPS were measured using a K-Alpha+ XPS spectrometer (ThermoFisher Scientific, East Grinstead, UK). The data were acquired and processed using the Thermo Avantage software. All freeze-dried samples were analyzed using a micro-focused, monochromated Al K_α X-ray source (400 μm spot size). The K-Alpha+ charge compensation system was employed during analysis, using electrons of 8 eV energy, and low-energy argon ions to prevent any localized charge build-up. The spectra were fitted with one or more Voigt profiles (BE uncertainty: ± 0.2 eV) and Scofield sensitivity factors were applied for quantification.[41] All spectra were referenced to the C 1s peak (C-C, C-H) at 285.0 eV binding energy controlled using the well-known photoelectron peaks of metallic Cu, Ag, and Au, respectively.

Isoelectric points

The TiO_2 suspensions (2 g/L) were freshly prepared by sonicating the corresponding nanoparticulate powder in UPW until a homogeneous suspension was obtained. The suspensions were then diluted and added to 40 mL of a 10 mM NaCl (p.a., Roth, Germany) solution to reach a final concentration of 0.1 mg/L. The pH of the suspensions was then adjusted (C863, Consort, Belgium) using HCl (Rotipuran, Roth, Germany) and NaOH (ISO p.a., Roth, Germany) to the target values. The suspensions were transferred to 3 mL polystyrene cuvettes and measured immediately for ζ-potential. The isoelectric points determination was triplicated.

ζ-potentiometry

Since determining the ζ-potential of the exposed suspension directly after sampling was not possible for logistical reasons, the freeze-dried exposed particles were resuspended in water sampling at the exposure site and kept frozen till the ζ-potential determination. The freeze-

drying and resuspension procedures did not affect the measured ζ-potential for P25 particles exposed to fulvic acids in a batch experiment (see supporting information for details).

An aliquot of 2.2 mL of surface water was transferred to Eppendorf tubes and centrifuged at 5 000 rpm (4 050 g) for 60 min (Universal 320, E4123 rotor, Hettich, Bäch, Switzerland). Test-experiments have shown that even for the most turbid samples, this step results in a clear suspension with a scattering signal too low to contribute to the ζ-potential determination. 1.5 mL of supernatant was sampled carefully from the top. The supernatant was transferred into a polystyrene cuvette and 1.5 µL of nanoparticle suspension were added to reach a target concentration of around 0.5 mg/L and vortexed for some seconds. This added suspension was produced by dispersing the corresponding freeze-dried nanoparticles in the centrifuged surface water (500 mg/L) using an ultrasound bath for 30 min. The cuvette was then transferred to the ζ-potential instrument (Wallis, Cordouan Technologies, Pessac, France) for measurement using laser Doppler velocimetry. The measurements were carried out at 25°C and the ζ-potential was calculated from the electrophoretic mobility using Smoluchowski's approximation. 10 replicates were measured per sample. A calibration of the instrument was required for optimal accuracy and was carried out with the standard reference material 1993 from NIST. The ζ-potential of the certified reference material measured again after calibration was -56 ± 4 mV (standard deviation n = 6) and matched the certified reference values of -58 ± 5 mV.

Thermogravimetry quadrupole mass-spectrometry (TG-QMS)

The thermogravimetric and mass-spectrometric analyses were performed with STA 449 F3 Jupiter® simultaneous thermal analyser fitted with type-S thermocouple (Pt-Pt/Rh) DSC/TG Octo sample carrier (Netzsch-Gerätebau GmbH, Selb, Germany) and coupled with the QMS 403 Aeolos® Quadro quadrupole mass spectrometer (Netzsch-Gerätebau GmbH, Selb, Germany) via a heated transfer line at 300 °C, untreated fused silica capillary, $l = 2.2$ m, $d = 75$ µm (SGE Analytical Science, Ringwood, Victoria, Australia). The freeze-dried

nanoparticles exposed to surface water were weighted in $PtRh/Al_2O_3$ crucibles using the Cubis® II Ultra-Micro lab balance (Sartorius AG, Göttingen, Germany). The sample mass was approx. 10 mg. An empty $PtRh/Al_2O_3$ crucible was used as a reference. The samples were heated from 45 °C to 100 °C (heating rate 10 °C min^{-1}), held at 100 °C for 15 min (isothermal step to remove moisture), and then heated from 100 - 1000 °C (heating rate 10 °C min^{-1}). Heating took place in an oxidative atmosphere of 50 mL min^{-1} synthetic air (N_2/O_2, 80/20 %) and 20 mL min^{-1} argon as protective gas. The combustion gases produced were ionized by the ion source (electron ionization) and recorded at the detector (secondary electron multiplier). The ion detection measurement was performed with selected mass-to-charge ratios (m/z) for detection at m/z 18 (H_2O) and m/z 44 (CO_2). Data evaluation was done by the Software NETZSCH Proteus Thermal Analysis (Netzsch-Gerätebau GmbH, Selb, Germany). The percentage mass losses of organic molecules and carbonates in the thermogravimetric curves were determined based on the combustion/decomposition temperatures and the QMS signal m/z 44.

Data analysis

Data exploration and preparation

The preparation of the raw data was carried out using Excel (Version 2309) and further analyses and data visualization with R-Studio (version 2024.12.0, Build 467). All features were numerical, except for “Season” and “Part” (particle type), which were converted to factor. Random Forest and XGBoost can deal with categorical variables in contrast to regression models. Therefore, the two categorical features were not included in the elastic net, partial least square regression, and linear models. This could be justified by the low importance of these features as determined in Random Forest and XGBoost models and the lack of any visible dependency as observed graphically (see discussion). Nonetheless, the absence of these two features could result in a slight disadvantage for the regression models.

A correlation map and scatterplots (Spearman's method, corrr package) were used to spot strong correlations between the feature variables as they should be removed for better model performances.[42] A strong correlation between [Na^+] and [Cl^-] existed over the whole concentrations' range and, therefore, we merged these two variables by summation. Correlations between some of the fluorescence indices lead to the exclusion of "Fluo_m" and "Fluo_c" to avoid redundancies. We plotted density distributions to check that the range of possible values was well covered by our dataset and carried out MDS analysis on the numerical variables to check that the data points were well distributed over the variables' space. By inspecting the scatterplots, we noticed that several important variables such as [Mg^{2+}] and [Ca^{2+}] have a logarithmic dependency with the ζ-potential. Therefore, we tested the regression models with and without asinh or log-transformation of the variables showing such non-linear dependency. As we observed slightly better performances with the transformed data, we show and discuss only those results here. The asinh function was used for elastic net and partial least square models instead of the log function to avoid issues for values close to or equal to zero. Furthermore, we split the dataset by particle type (P25, NM103, etc.) and tested Random Forest and Elastic net models on each of the five datasets. We obtained similar but less stable results than with the merged dataset with the particle type as an additional categorical variable. Therefore, we describe and discuss only the results for the merged dataset here.

Models training

All models were trained following the same procedure. At first 50 random data partitions were generated with 80% of the data defined as the training set and 20% as the test set. The model was then trained and optimized using the training set to minimize the root mean square error (RMSE). The physicochemical parameters of the surface waters were taken as input features and the ζ-potential as output variable. We determined the RMSE of the predicted

values for the test set using the model with optimal parameters. Thus, we obtained 50 models with their corresponding RMSE for each evaluated method. We tested univariate linear regression, elastic net regression, partial least square regression (PLSR), random forest, and extreme gradient boosting (XGBoost). While there is no universal rule for the minimum number of required data points, our evaluation of model stability, performance metrics, and variable importance distributions indicates that the sample size used in this study is sufficient for reliable analysis.

A simplistic one-dimensional linear regression model, denoted as "linear regression" in the following text, was tested along with the multivariate models. Here only log([Mg^{2+}]) was used as the input variable as it is the most important variable in both XGBoost and Random Forest models. We also trained elastic net regression models, which are L_1 and L_2 regularized multivariate linear regression models,[42] on the scaled data with the caret and glmnet packages using cross-validation. PLSR models are reduced rank regression models finding a linear regression model by projecting the predicted variables and the observable variables to a new space of maximum covariance.[42] We built and optimized PLSR models using the pls and caret packages, respectively. The optimal number of components to be included in the model was selected based on the smallest RMSE determined by cross-validation.[42] Random forest models are an ensemble of decision trees trained on a small set of input features and were well performing with similar datasets.[32] We optimized the parameters "mtry", the sample size, and the node size based on the out-of-bag error using a hypergrid (R-package "Ranger"). The number of trees was set to ten times the number of input variables. Finally, XGBoost models are ensembles of regularized Regression trees optimized using a gradient boosting approach.[43] We built and trained XGBoost models using the xgboost package. The maximal depth and the number of iterations were optimized based on the RMSE of the test set.

Model exploration

To explore the XGBoost model and connect its functioning with physico-chemical processes known from the literature, we investigated a random selection of models using importance values, Shapley values (both using the xgboost package), partial dependence and individual conditional expectation (ICE) plots (pdp package), and surrogate decision trees (packages rpart and rpart.plot). We used the gain parameter to quantify the importance values of the input features. Gain represents the improvement in accuracy (loss reduction) brought by a feature to the regression trees' branches it is used in.[43] Therefore, it measures how much each feature contributes to the model's predictive power, with higher values indicating more important features. We used the importance values to rank the input parameters by their influence on the ζ-potentials. The Shapley (shap) values are the natural logarithm of the odds shift induced by an input feature for a given data point.[44] In other words, a high positive Shapley value indicates that the model's output value for the given data point is very likely to be strongly shifted towards higher values due to the considered feature's value. In contrast, negative Shapley values indicate that an increase in the feature value will result in a higher probability to obtain a lower output value. We used Shapley values to further explore the importance of variable on a single data point basis. ICE plots show how the model's individual predictions change when varying a single input feature, while Partial Dependence plots display the average effect of a feature on model predictions across all instances.[45] These plots provided us a more intuitive understanding on how the model respond to changes in the most important variables, reflecting underlying relationships between input and output variables. We built surrogate decision trees for the XGBoost model as decision trees are more intuitive to understand and interpret. For that, we replaced the measured output values with the ones predicted by the XGBoost model and trained a decision tree on this dataset. We selected the optimal cp parameter (pruning) to minimize prediction errors.[42] For more details, please consult the complete R-scripts and raw data available at https://zenodo.org/records/14965814.

Results and discussion

Sunscreen extract characterization

The mean and median Feret diameters determined from the SEM images (n = 306) are 72 nm and 68 nm, respectively, which is slightly larger than the sizes found for nanoparticles extracted from other sunscreens.[27] The mean and median aspect ratios determined from the SEM images are 1.56 and 1.42, respectively. The particles' morphology was quite irregular (Figure 1). The particles were hydrophobic as they were difficult to suspend in water.

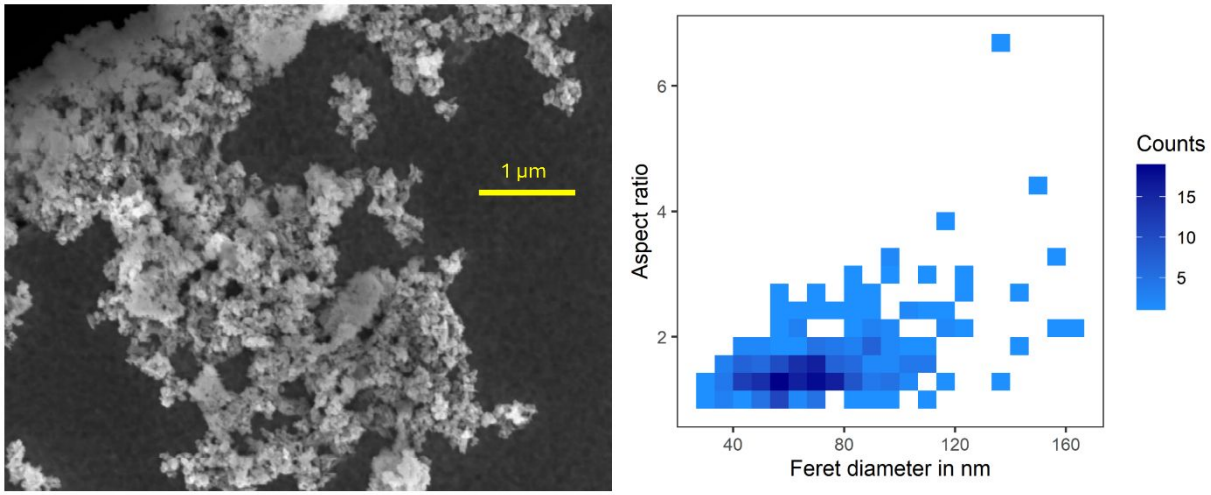


Figure 2: representative SEM image (left) and size and aspect ratio distribution (n = 306) of the TiO_2-nanoparticles extracted from the commercial sunscreen and used in this study (right).

The isolectric point for the R-TC90 and the paint extract were close with 5.5 and 5, respectively, confirming that both types of particles have strong similarities (Figure S3). P25 had ζ-potentials close to zero at pH 5-7, in accordance with previous measurements and despite lower ζ-potential magnitudes than previously reported.[46] This discrepancy could be explained by the presence of aliphatic impurities accumulated during storage and detected using XPS (see below). NM-103 and the sunscreen extracts were neutral above pH 6 with slightly positive ζ-potentials below which is consistent with a PDMS coating.[27]

Data exploration

Overall, the input parameters cover a wide range of values which could be found in surface waters demonstrating the representativity of our exposure sites (Figure 3). The MDS ordination diagram (Figure S1) shows that the data points are well distributed over the projected space indicating that the diversity in the dataset is high as required for an accurate and generalizable prediction. The mean, median, mode, and standard deviation of the ζ-potentials measured in our samples were -14.5, -13.0, -20, and 14.5 mV, respectively. The density distribution shows that many values are close to zero (Figure 3). Therefore, negative ζ-potentials are much more common under environmental conditions than positive ones and the absolute values are relatively low. This agrees with values reported for seven natural and reconstituted waters.[47] Therefore, we do not expect a high electrostatic stabilization for most of the naturally coated TiO_2-particles.

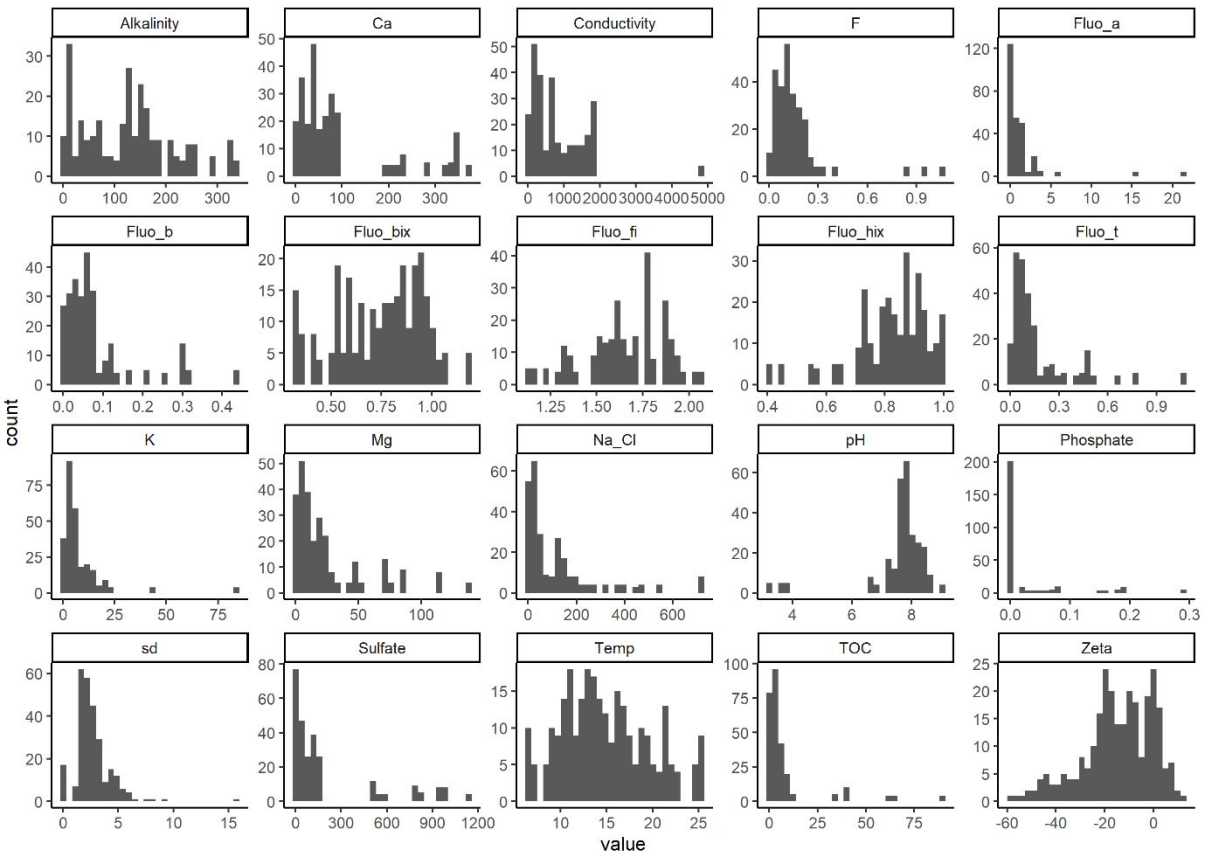


Figure 3: density distribution (n = 269) of all numerical parameters in the dataset used in this study. The parameters are sorted in alphabetical order. Ca, F, K, Mg, Phosphate, and Sulfate denote the concentration of the corresponding ions in mg/L. Na_Cl is the sum of Na^+ and Cl^- concentrations in mg/L. Conductivity is given in µS cm^{-1}, total dissolved organic carbon (TOC) in mg/L, ζ-potential and the standard deviations (sd) of the individual ζ-potential measurements (n = 10) in mV, and average water temperature during the exposure in °C. The parameters named Fluo_X are unitless descriptors of the organic matter composition obtained from the fluorescence maps as explained in the method description.

There is no visible correlation between the ζ-potential and the particle type or season of exposure (Figure 4). This suggests that these categorical variables are not important for predicting ζ-potentials. We could not find any correlation between season and water composition; therefore, it is not expected to influence the surface charge. However, the lack of trend with the particle type is surprising as the surface properties of these particles are different. For instance, P25 are composed of bare TiO_2 whereas the other particles had an Al or PDMS based coating. The poor stability of the hydrophobic coating in aqueous suspension as observed in laboratory conditions with nanoparticles used in sunscreen, for example, [26,48] could explain this observation. The coverage with DOM could also mask the underlying coating as observed for other anthropogenic particles.[21,49] In both cases, the ζ-potential could be dominated by the water chemistry rather than the surface chemistry.

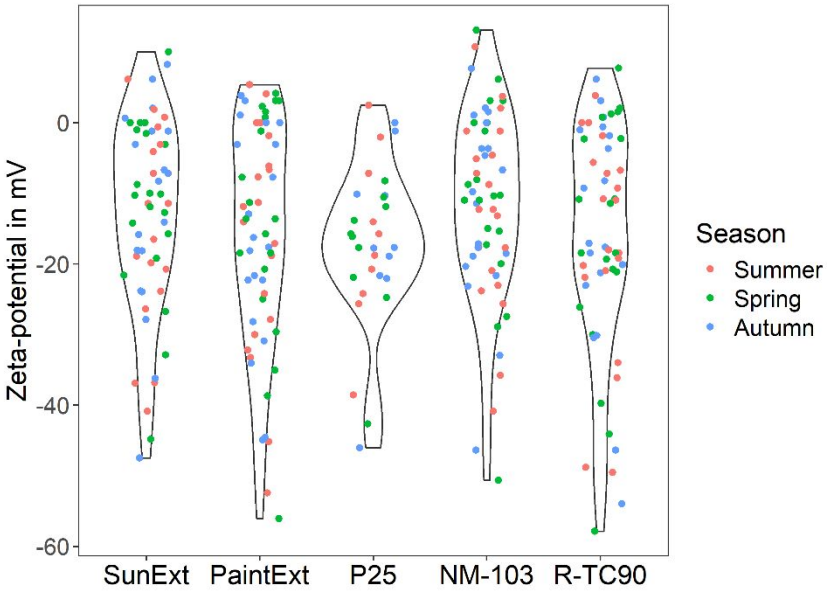


Figure 4: distribution of measured ζ-potential values (n = 269) as a function of particles type (x-axis) and Season (color).

Selection of the best modelling method

The XGBoost models have on average the best RMSE for predicting the ζ-potential from the water physico-chemical parameters (Figure 5). Furthermore, these models run much faster than random forest models thanks to more advanced and optimized algorithms for training the regression trees.[43] Further, it has been shown that XGBoost outperforms neuronal networks or linear models for tabular datasets from medicinal research.[44] The best RMSE is close to the median standard deviation of the ζ-potential measurements determined from 10 replicates, which is around 2.4 mV (Figure 3). This suggests that the performance of the XGBoost models is close to the best achievable error rate considering measurement uncertainties. The elastic net and PLSR models have slightly lower RMSE than the simplistic univariate linear model, which has a median RMSE of 6 mV. Therefore, using $[Mg^{2+}]$ alone as an input feature could be reasonable for applications in which a rough estimation is sufficient as only minor improvements were reached by adding other variables. The tree ensemble methods, however, clearly outperform the linear models. This indicates that multivariate non-linear effects are

affecting the ζ-potential and, therefore, models which can handle complex non-linear interactions should be preferred.

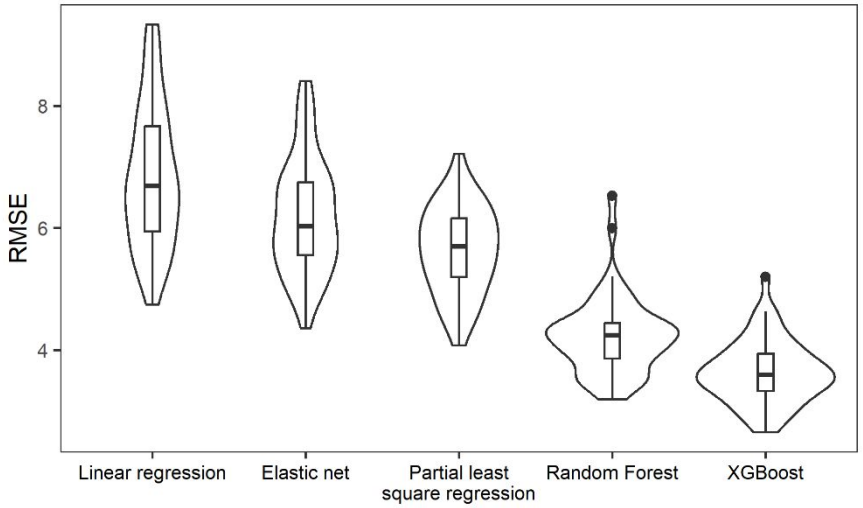


Figure 5: model performances comparison (root mean square error on the test dataset, RMSE) of the models tested to predict ζ-potential. These values are much lower than the standard deviation of the ζ-potentials in the whole dataset was 14.5 mV indicating the mean error corresponding to a random guessing process. The distributions represent the RMSE of 50 models generated from different random training/test data-partitions.

A low RMSE for the prediction and an excellent fitting between prediction and measurements (Figure 6) suggests that the XGBoost models are also informative as they captured the complex relationships between the variables. Indeed, it has been shown that more complex models with low RMSE are a better choice for interpretation than simpler but less accurate models.[44] Therefore, we recommend using XGBoost for predicting ζ-potentials in surface waters and exploring these models to understand the relationships between environmental variables and ζ-potential.

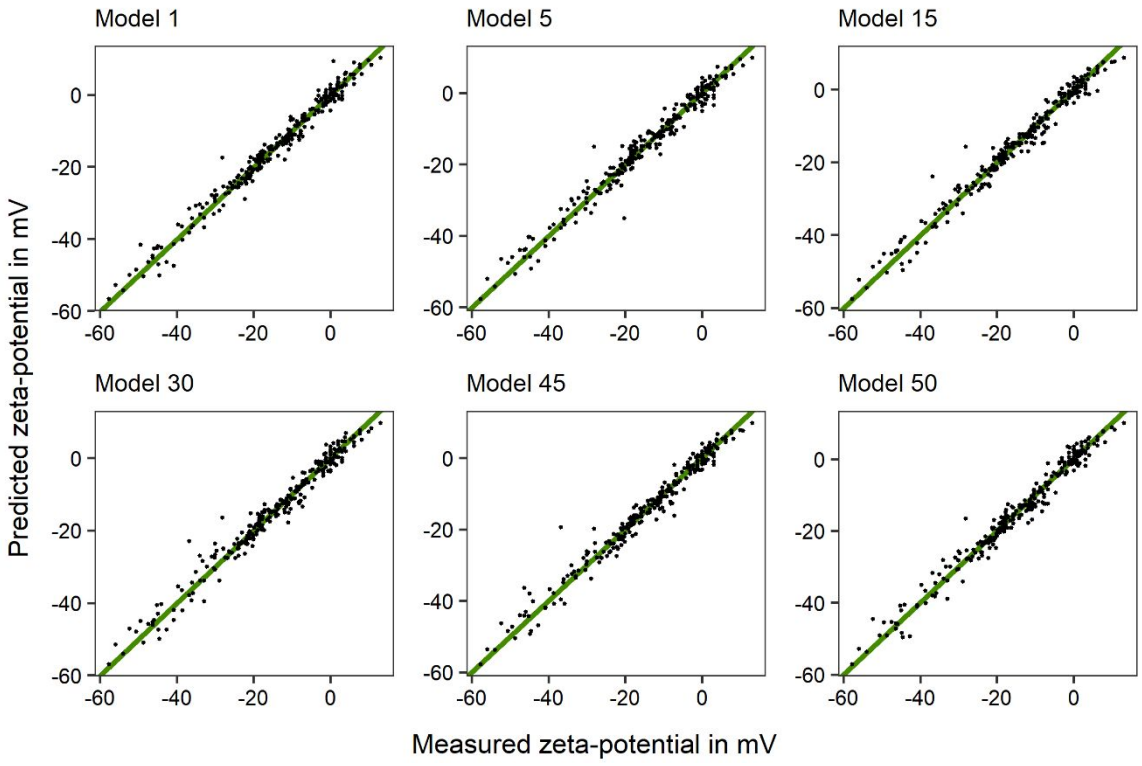


Figure 6: ζ-potentials values predicted by the XGBoost model versus the measured values for a selection of six among 50 randomly generated models. The green line denotes the y = x function. All datapoints were used for these plots. The different models are consistent in terms of accuracy. The very good alignment of the data points with the y = x lines illustrates the excellent prediction power as indicated by the low RMSEs.

Model exploration

Among the input variables, [Mg^{2+}] has the highest importance for the XGBoost models' prediction accuracy (Figure 7), followed by [Ca^{2+}]. Divalent cations are known to shield the surface charge of colloids better than monovalent cations (Shulze-Hardy rule).[50–53] Further, divalent cations can form complexes with DOM and the hydroxylic groups on the surface of metal-oxides more easily than monovalent cations.[29,53–55] Divalent anions, in particular sulfate, were not as effective for charge destabilization as cations for other types of nanoparticles.[56,57] Therefore, the high importance of the divalent cations' concentration for predicting the ζ-potential of TiO_2-particles was expected. However, it is interesting that the

$[Mg^{2+}]$ has a much higher importance than $[Ca^{2+}]$ since the median $[Mg^{2+}]$ was around 4.5 times smaller than the median $[Ca^{2+}]$. Although Mg^{2+} had a stronger colloidal destabilization effect on hematite nanoparticles than Ca^{2+},[57] the effect on the electrophoretic mobility of illite soil colloids[58] and goethite particles as well as natural particles isolated from river[52] were comparable for both ions. Therefore, it is unlikely that a higher affinity of Mg^{2+} would result in a much stronger influence on the ζ-potential despite the much lower concentration. The XPS analysis of particles exposed to a selection of sites confirms this statement as the ratio of the atomic % of Mg/Ca detected at the particles' surface did not show any trend following either the ζ-potentials nor the $[Mg^{2+}]/[Ca^{2+}]$ in the solution (Figure S4). Therefore, although a stronger difference in affinities for Mg^{2+} compared to Ca^{2+} was observed for NM-103 compared to other particles, it seems unlikely that an effect of this magnitude could explain the strong influence of $[Mg^{2+}]$ on the ζ-potential for all particle types. The strong correlation between $[Mg^{2+}]$ and $[Ca^{2+}]$ (Pearson correlation coefficient = 0.934) is a more plausible explanation for the large difference in importance. Indeed, the linear fitting of the ζ-potential against asinh($[Mg^{2+}]$) (R^2 = 0.7652) is only slightly better than for asinh($[Ca^{2+}]$) (R^2=0.7386), suggesting that the prediction potential of both parameters is comparable. The XGBoost algorithm would in such cases select $[Mg^{2+}]$ as the best split for most of the random partitions and neglect $[Ca^{2+}]$ because it is redundant with $[Mg^{2+}]$ due to the high correlation. Therefore, the higher performance of $[Mg^{2+}]$ could be due to subtle differences in the noise in the dataset and does not necessarily reflect a physicochemical effect. It should be noted that random forest models were trained using input variable sets in which calcium and magnesium concentrations were treated both as separate variables and as a combined variable. In both approaches, the resulting models exhibited similar root mean square error values and comparable variable importance rankings (Figure S5). Consequently, we concluded that merging the two variables does not significantly alter the fundamental characteristics of the model. Nevertheless, to obtain more detailed information regarding the individual

contributions of calcium and magnesium, we decided to retain the models in which these variables were not combined.

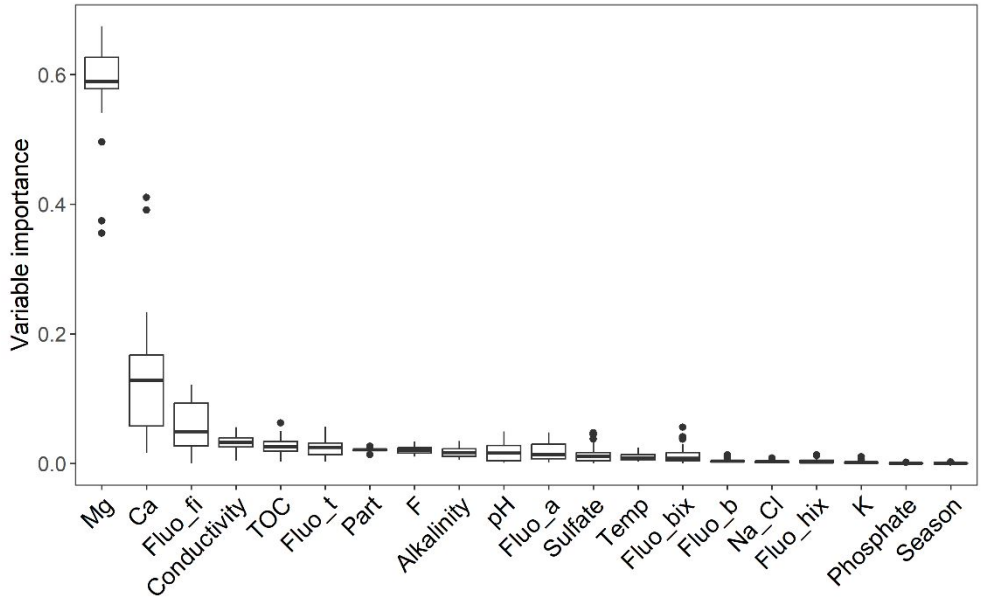


Figure 7: variables importance of the input parameters (water composition, particle type, temperature, and season) for the accuracy of the prediction of ζ-potentials by the 50 randomly generated XGBoost models. Ca, F, K, Mg, Phosphate, and Sulfate: concentration of the corresponding ions. Na_Cl: sum of Na^+ and Cl^- concentrations. TOC: total dissolved organic carbon (TOC), Temp: average water temperature during the exposure. The parameters named Fluo_X are descriptors of the organic matter composition obtained from the fluorescence maps as explained in the method description. Part: particle type.

The ICE plots for [Mg^{2+}] reveal an asymptotic behavior of the ζ-potential prediction as a function of the cation's concentration consistent among the models (Figure 8). This behavior corresponds well with the log-shape of the ζ-potential versus [Mg^{2+}] as indicated by the low RMSE of the univariate linear models ($\zeta = a * ln([Mg^{2+}]) + b$, with a and b the fitting parameters). This agrees with laboratory results showing an asymptotic behavior of ζ-potential with increasing ionic strength.[59,60] In a simple suspension, once all surface charges are shielded by the cations, the ζ-potential tends toward zero and additional ions would not affect it. However, we observed positive predicted ζ-potentials for high [Mg^{2+}]. This has been

observed under laboratory conditions as well and is explained by the sorption of cations.[59,61] Under environmental conditions, such sorption could also occur through DOM bridging[29] or as surface precipitation of $CaCO_3$[32] which could influence the surface charge. However, it is unclear to which extent such precipitates could influence the ζ-potential as the solubility of $CaCO_3$ complicates the experimental determination of ζ-potentials in its pure form.[25]

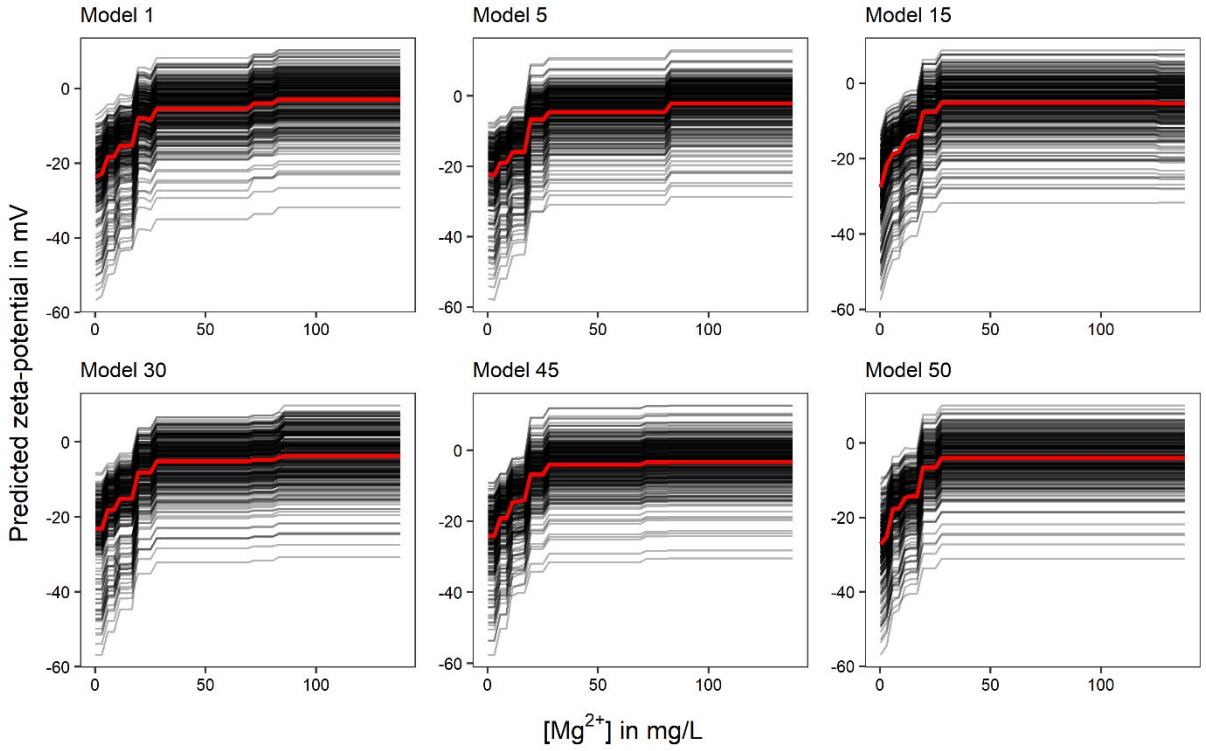


Figure 8: partial dependence (red lines) and individual conditional expectation (grey lines) plots of the input variable $[Mg^{2+}]$ for a selection of six among 50 randomly generated XGBoost models.

The behavior revealed in the ICE graphs is consistent with the shap values' distributions for $[Mg^{2+}]$ (Figure 9). Indeed, $[Mg^{2+}]$ follows the shap value: when the concentration is small, the predicted ζ-potential will decrease and vice versa. The cluster of the large concentrations around a shap value of 10 is consistent with the asymptotic behavior of the corresponding ICE curves. These very large absolute shap values reflect the dominating effect of the $[Mg^{2+}]$ on the ζ-potential.

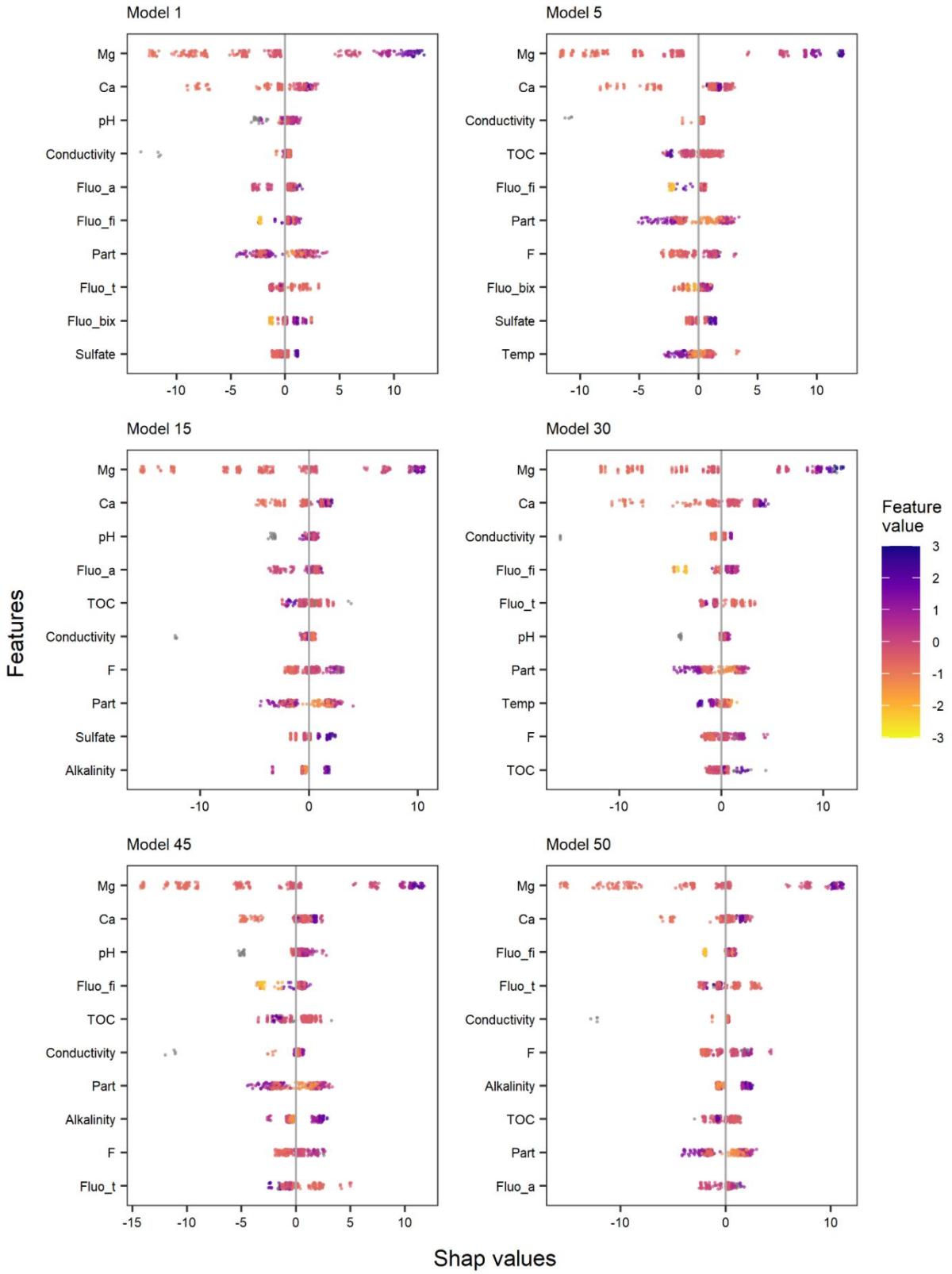


Figure 9: Shapley (Shap) values distributions (x-axis) for the ten most important variables (y-axis) for a selection of six among 50 randomly generated XGBoost models. The Shap values are the natural logarithm of the odds shift of the prediction due to the evaluated variable. A high negative or positive

value denotes a high probability that the predicted output value will be low or high because of the input variable. The input variables are sorted by importance. The color legend denotes the normalized values of the input variables, the darker the higher the value. For the particle type ("Part"), a numerical factor from 1 to 5 was assigned to the five types of particles in the following order: sunscreen extract, paint extract, P25, NM-103, and R-TC90.

The effect of $[Ca^{2+}]$ as observed in the ICE plot and shap values distribution is similar to $[Mg^{2+}]$ (Figure S6). This confirms our interpretation of the difference in importance between $[Mg^{2+}]$ and $[Ca^{2+}]$. These two divalent cations most probably reflect the same underlying physicochemical effects and are entangled in the model due to their high correlation with each other.

Interestingly, the fluorescence index (Fluo_fi) was the third most important parameter after the divalent cations' concentrations (Figure 7). This index indicates if the precursor material for DOM is rather microbial (FI ≈ 1.8) or terrestrially derived (FI ≈ 1.2).[39] However, such interpretations are correlative by nature and should be taken with care. Nonetheless, we can assume in the context of this study that the fluorescence index is a proxy for DOM quality. On the other side, dissolved organic carbon concentration (TOC) is the fourth most important variable, with much lower importance than the fluorescence index. Considering that the fluorescence index does not correlate with TOC (Pearson correlation coefficient = -0.591), we can conclude that DOM quality has more influence on the ζ-potential than DOM quantity. We explain this observation by the fact that TOC values are never extremely low in most natural waters as reflected in our dataset (min(TOC) = 0.2 mg/L). In other words, there could be enough organic molecules to saturate the nanoparticles' surface, even in surface waters poor in DOM, for example, through irreversible sorption. Indeed, in the dialysis bag setup, solutes—and, by extension, dissolved organic matter (DOM)—are continuously supplied to the membrane surface due to convections and advection, resulting in a sustained excess of

DOM in the dialysis bag. To determine an order of magnitude for the quantity of organic matter in the environmental coating, as an example, we determine the % of organic C using TG-QMS for NM-103 exposed to sites 15 and 61 (exposed during fall and summer, respectively). Site 15, a lake connected to the Rhine River near Speyer, Germany, exhibits a composition typical of most sites in our dataset, without any notable distinguishing features. Site 61 was from a wetland near Bremen, Germany. It was, therefore, rich in humic substances and had the highest TOC in the dataset. Based on TG-QMS measurements, we observed a clear increase in the relative mass of organic molecules for both sites compared to the pristine NM-103 particles (0.9%)[34] with 2.8% for site 15 and 7.2% for site 61 (Figure S7). This suggests that for most sites, there was a significant amount of organic matter sorbed on the surface. Therefore, a low amount of organic molecules sorbing on the surface cannot explain the low influence of TOC on the ζ-potential. Nonetheless, high TOC induces lower (more negative) ζ-potential for four of the explored models (Figure 9) which agrees with reports of ζ-potential determined in the presence and absence of DOM.[46,47,59,60,62,63]

The effect of DOM quality on the ζ-potential of TiO_2-particles has been observed in lab experiments comparing DOM with different aromaticity[62] or molecular weight,[64] protein or humic acid content,[65] or DOM from terrestrial or aquatic origin.[63] Our results obtained under environmental conditions are consistent with these observations. The ICE curves reveal a more complex but consistent effect of the fluorescence index on the influence on the ζ-potential (Figure 10). Two thresholds can be observed at values around 1.35 and 2 corresponding to strong terrestrial or microbial origins, respectively. In average, the ζ-potential is predicted to be slightly higher (less negative or more positive) between these two values. This could be interpreted in the light of results showing that proteins can compete with humic acids for the sorption sites on TiO_2.[65] Such effects could be relevant under environmental conditions only for extreme cases in which the DOM results mainly from

microbial production with a high level of polysaccharides and proteins or, on the contrary, where DOM has a high amount of humic substances (e.g. from peatlands). Although consistent over the sampled models, the effect of the DOM quality remains secondary compared to the effect of divalent cations. The shap values distributions support this interpretation by revealing that the predictions are strongly influenced (shap ≈ 4) for only a subset of the datapoints with the lowest fluorescence indexes of the dataset (Figure 9).

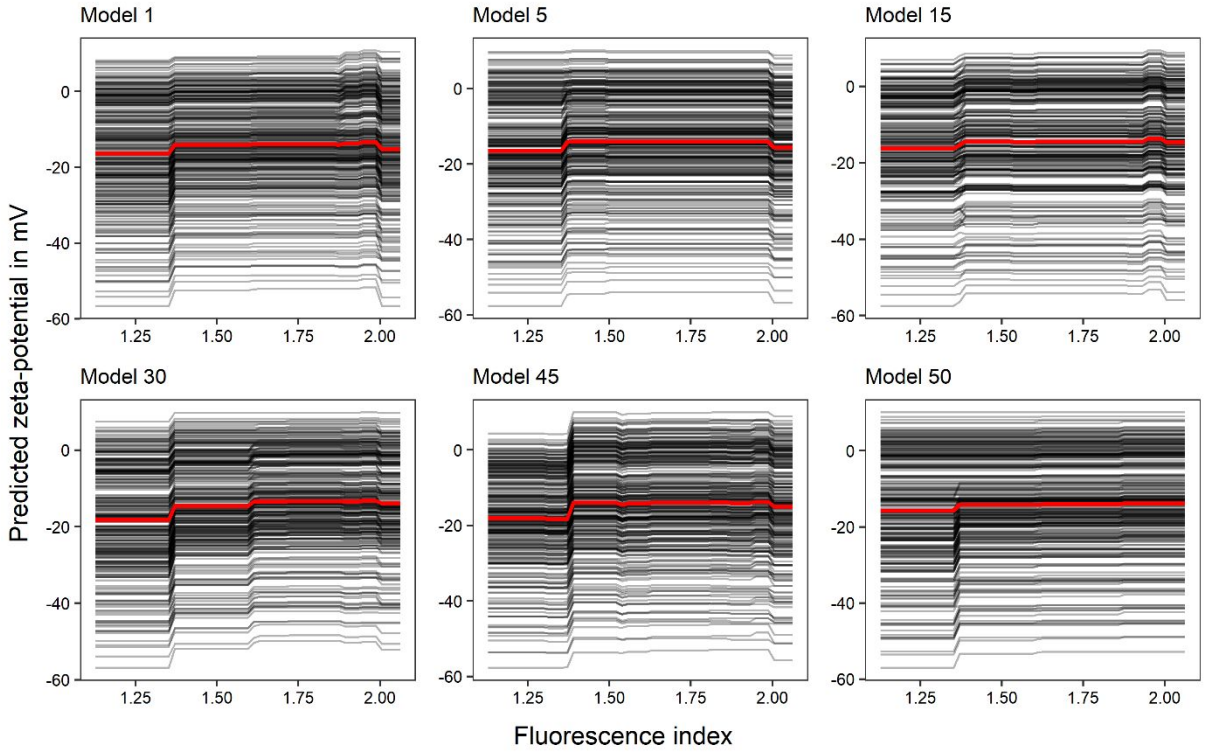


Figure 10: partial dependence (red lines) and individual conditional expectation (grey lines) plots of the fluorescence index for a selection of six among 50 randomly generated XGBoost models.

Surprisingly, the importance of pH is low (10th higher importance, Figure 7). The protonation of hydroxyl groups on the surface of metal oxides depends directly on the pH of the suspension[66], and, thereafter, pH has been observed to be a critical parameter for the ζ-potential of TiO_2-particles in simplified matrices.[47,59,61] The shap distributions revealed that for the models including pH (models 1, 15, 30, and 45), the parameter was only influential for a subsample of the data points with extreme values (Figure 9). This is not consistent with the

literature in which pH has been shown to have a high effect on the surface charge of TiO_2-particles.[59,67,68] We explain this discrepancy by the low range of pH values in our dataset. Indeed, in batch experiments, extreme pH values are usually set in addition to neutral values to explore effects over the whole possible range of pH in the environment. In most surface water, however, the pH never reaches such extreme values. Furthermore, the nanoparticles used in this study pH had few effects on the ζ-potentials in the pH range of 5-8 (Figure S3), most probably due to the hydrophobic coatings. Therefore, it is possible that the effect of pH is not noticeable under typical environmental conditions but would become important in special cases such as waters draining peatlands.

Similarly, conductivity (4th rank in importance, Figure 7) is important for only a small subset of data points with extreme values (Figure 9). Conductivity can be considered as a proxy for ionic strength, which is acknowledged as a key parameter for ζ-potential.[50] Since the conductivity correlates moderately with [Mg^{2+}] (Pearson correlation coefficient: 0.632), we can conclude that the ionic strength can bring additional information and, hence, become important instead or together with divalent cations concentration under specific conditions.

Intuitive understanding of the models

Surrogate decision trees are convenient to interpret complex multivariate models intuitively. In this regard, we generated decisions trees approximating the predictions of the six explored XGBoost models (Figure 11). For all surrogate models, [Mg^{2+}] consistently dominates the decision process, but many parameters are involved in each tree. This complexity suggests that an accurate prediction of the ζ-potential requires many secondary parameters, for instance, the particle type, sulfate concentration, pH, alkalinity, and TOC. Furthermore, the tree representation makes it clearer that the paint extract and R-TC90 are treated differently from the other particles in the decision process, which we explain by the similar characteristics of these pigments used in paints.

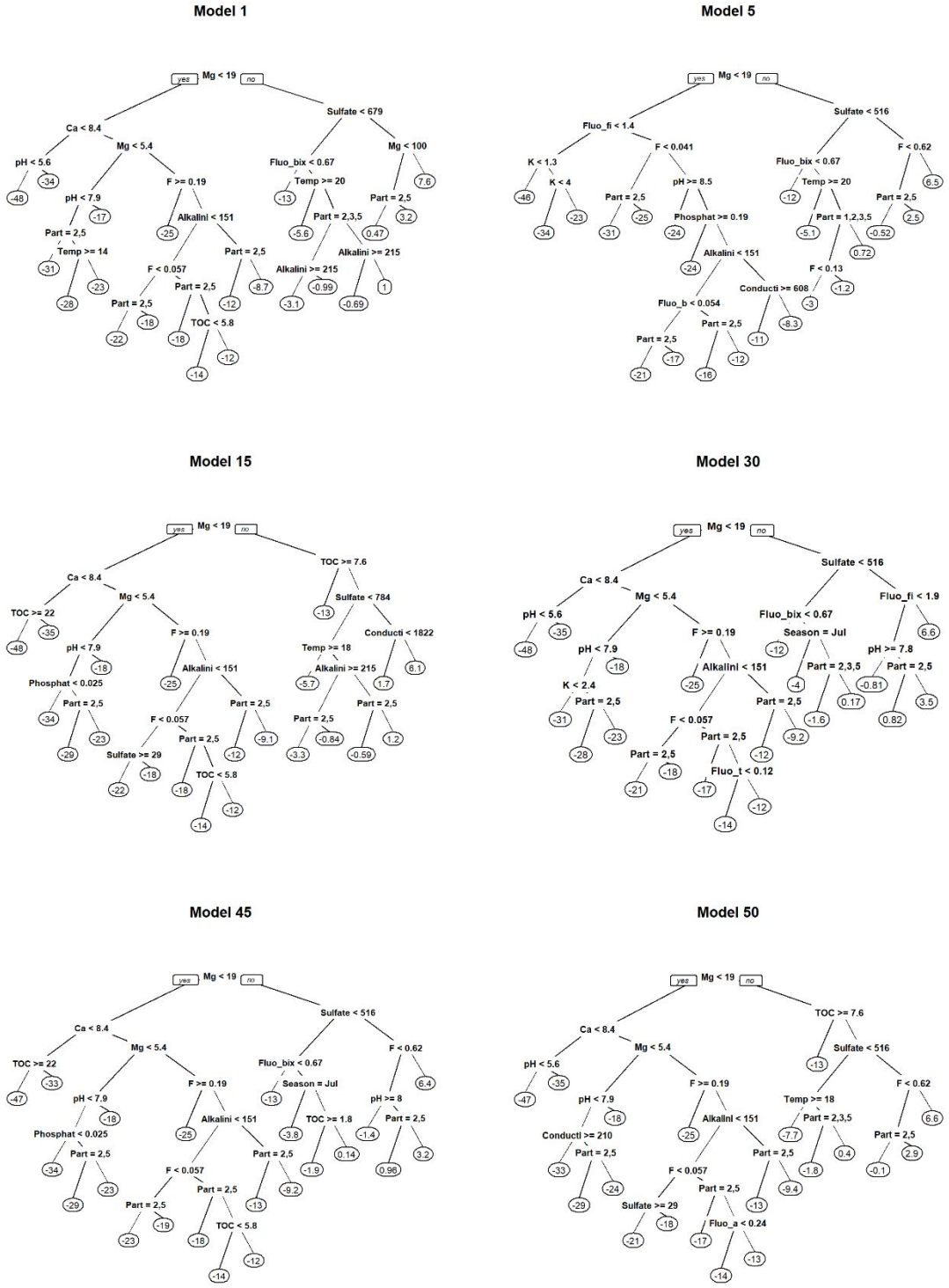


Figure 11: surrogate decision trees generated for a selection of six among 50 randomly generated XGBoost models. The leaves of the trees are the predicted ζ-potentials and the numbers at the nodes next to the input variables' names denote the threshold values for the split decision. For the categorical

variable particle type ("Part"), a numerical factor from 1 to 5 was assigned to the five types of particles in the following order: sunscreen extract, paint extract, P25, NM-103, and R-TC90.

Connecting ζ-potential with environmental coating's composition

To explore the connection between ζ-potential and environmental surface coating, we investigated the surface composition of a selection of nanoparticles and sites presenting contrasting water compositions using XPS. Despite their differences, the sites showed similar increase in the proportion of C-O and O-C=O groups and a concomitant lower proportion of C-C/C-H compared to the non-exposed particles (Figure 12). Carbonate groups' content increased for a few samples and their proportions were lower than other carbon species. Despite the increased proportion of O–C=O functional groups—potential carriers of negative surface charge—on the exposed particle surfaces, no statistically significant correlation was observed between the normalized O–C=O atomic concentration and the ζ-potential ($p = 0.07$). Although distinct differences were evident among particle types, a consistent trend in carbon speciation was observed, characterized by an increase in oxygenated species, including carboxyl groups (figure 12). In contrast, the ζ-potentials of these exposed particles exhibited substantial site-dependent variability, with values ranging from a minimum of –54 mV at site 49 to a maximum of +8 mV at site 47. These results suggest that alterations in the composition of organic surface coatings cannot fully explain changes in surface charge. This finding is consistent with the relatively minor role of dissolved organic matter (DOM) composition and concentration in predicting ζ-potential, particularly when compared to the influence of divalent cation concentrations. This may be attributable to a generally similar surface composition, which appears to be governed more by the intrinsic surface chemistry of the particles than by variations in DOM composition. For example, it is plausible that only specific molecules are capable of adsorbing in appreciable quantities onto the nanoparticle surfaces, and that the majority of these molecules are ubiquitous across different surface

waters. It should be noted, however, that while the composition of the sorbed organic layer may not be the primary determinant of surface charge, the quantity of DOM adsorbed onto the particles may still be relevant. Unfortunately, this parameter could not be quantified from the XPS data.

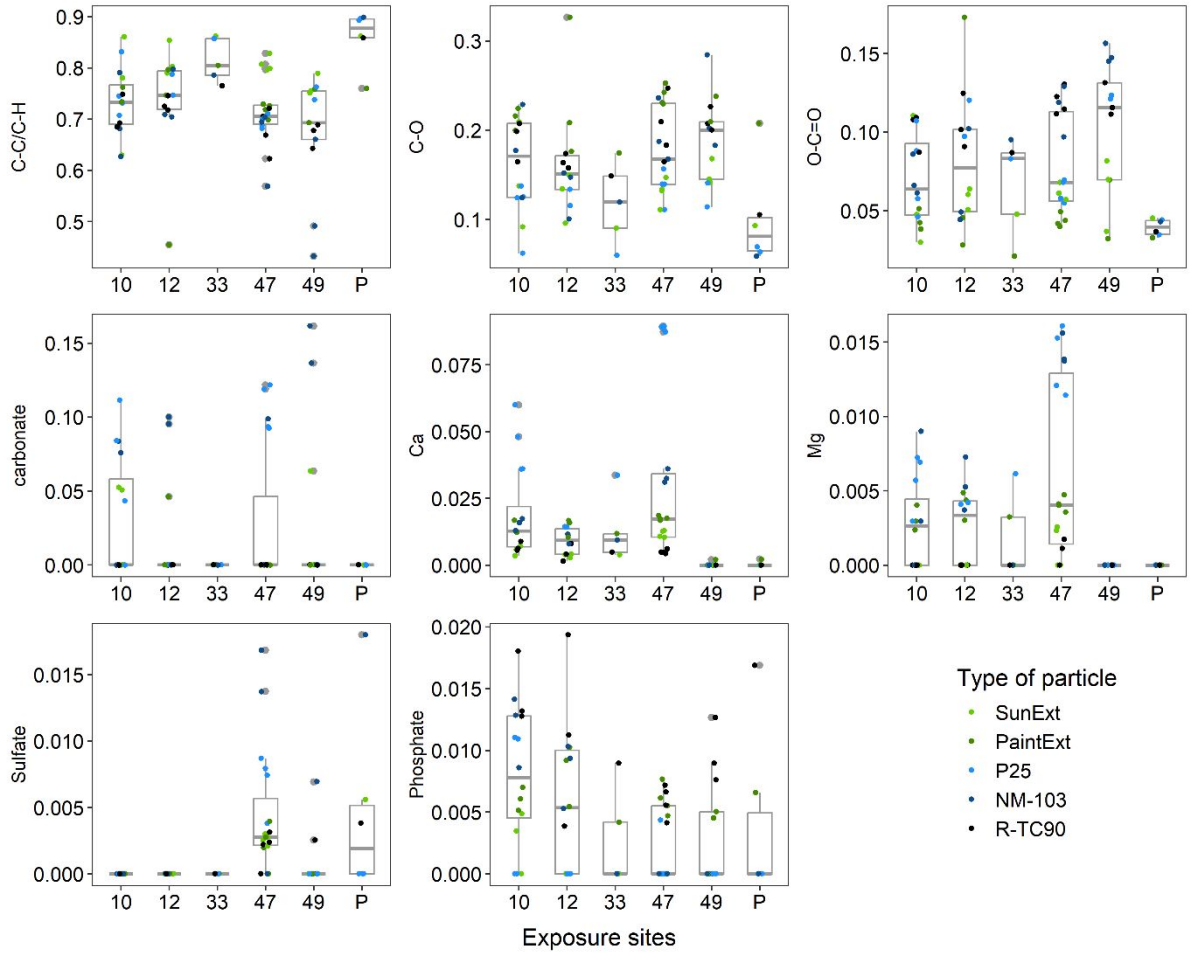


Figure 12: fractional contribution of each element detected on the TiO_2-nanoparticles before (P) and after exposure (numbers) to a selection of contrasting surface waters. The ratios are calculated based on the XPS quantitative analysis. For better comparability, all concentrations (at%) are normalized by the total carbon content concentrations of the elements in atomic %. See text for the surface waters' description.

The proportion of Ca, Mg, sulfate, and phosphate changed notably for some samples and sites, hence confirming that these ions are part of TiO_2-particles' environmental coating

(Figure 12). However, the final proportion remained very small compared to the C-signals. Thereafter, the sorption of these ions is most probably not influencing significantly the surface charge. The effect of the divalent cations on the surface should, therefore, mostly be the result of electrostatic shielding instead of a neutralization of the surface charges. Nonetheless, it is important to note that there are strong differences in the proportions of inorganic ions in the coating depending on the exposure site and the particle type. This indicates that, although the surface composition of the exposed particles does not influence strongly the ζ-potential, it does influence the composition of the environmental coating and, thereafter, properties such as sorption coefficient, hydrophobicity, and surface reactivity.

Conclusion

The proposed XGBoost models can predict with high accuracy the ζ-potential of the most common TiO_2-particles in temperate surface waters from standard physico-chemical parameters. However, these models are probably more useful for predicting large numbers of values based on data from water composition databases, for instance. In that case, the most important parameters to consider can be selected based on our model's exploration. In case not enough data is available, a univariate linear model using [Mg^{2+}] could already give a reasonable estimate.

The strong asymptotic decrease in ζ-potential in the presence of divalent cations under field conditions are consistent with results reported in previously published laboratory experiments. In contrast, the low effect of DOM quantity and pH under field conditions was not expected based on the literature, most probably due to the limited range of these parameters measured in environmental water. Although many parameters are needed for an accurate prediction of ζ-potential, the divalent ions' concentrations and the DOM quality are by far the most important parameters. In most batch experiments, surrogates such as Suwannee River humic acids are

used for simulating the effect of DOM. Our results suggest that more diversity and complexity are needed in the DOM-surrogates to obtain more realistic conditions. Furthermore, the concentration of divalent cations should always be systematically investigated, if possible, using both Mg^{2+} and Ca^{2+} to distinguish their entangled effects.

Currently, our results are limited to surface waters comparable to the ones present in our dataset, which was limited to temperate climates. Therefore, it would be useful to carry out similar experiments in other climatic regions (e.g. tropical, polar). However, the method is limited by the temperature range tolerated by the dialysis membrane used to expose the nanoparticles to surface waters (5-37°C, based on the supplier's information). Membranes with better stability at low and high temperatures would be useful for harsher climates. Our approach could also profit from a more in-depth characterization of the DOM composition using, for instance, high resolution mass spectrometry.

Our approach could also be applied to other types of particulate emerging contaminants such as microplastic, or to natural particles such as clays or iron oxides. It could also be used to predict other nanoparticles' properties under environmental conditions such as coating composition or surface reactivity, for instance.

Conflict of interest

There is no conflict to declare.

Data availability statement

The raw-data for this article as well as the complete R-code used for the data analysis, the modeling, and the sites selection are available at Zenodo at https://zenodo.org/records/14965814.

Acknowledgements

This research was founded by the German Research Foundation (DFG, project number: 458047880, PH 277/4-1 and LE 3784/8-1). The authors would also like to thank the Joint Research Center of the European Commission for providing us with the NM103 sample, the local authorities and water bodies owners who authorized us to carry out the field experiment at diverse locations and provided us with crucial information, Karin Meyer for the characterization of ions concentrations and TOC, Emma Tätenmeyer for helping with fluorescence measurements, Marc Fritzsche for helping with the 3D-printing, Jochen Rauls for his help during the field experiments, and Mathilde Knott and Gabrielle E. Schaumann for the organizational support as group leaders.

Authors contribution

AP: conceptualization, resources, supervision, investigation, data curation, data analysis, visualization, writing. NTSK: investigation, data curation, review & editing. MG: investigation, review & editing, FH: investigation, review editing, VT: investigation, data analysis, review & editing, OL: investigation, resources, supervision, data curation, review & editing, SK: data analysis, review & editing, MJGA: investigation, review editing, VW: investigation, review & editing, EDL: investigation, review & editing

References

1 Heilgeist, Sekine, Sahin and Stewart, Finding Nano: Challenges Involved in Monitoring the Presence and Fate of Engineered Titanium Dioxide Nanoparticles in Aquatic Environments, *Water*, 2021, **13**, 734.

2 Karlsson, Álvarez-Asencio, Bordes, Larsson, Taylor and Steenari, Characterization of paint formulated using secondary TiO2 pigments recovered from waste paint, *Journal of Coatings Technology and Research*, 2019, **16**, 607–614.

3 Baalousha, Yang, Vance, Colman, McNeal, Xu, Blaszczak, Steele, Bernhardt and Hochella, Outdoor urban nanomaterials: the emergence of a new, integrated, and critical field of study, *Science of The Total Environment*, 2016, **557**, 740–753.

4 Baalousha, Wang, Nabi, Loosli, Valenca, Mohanty, Afrooz, Cantando and Aich, Stormwater green infrastructures retain high concentrations of TiO2 engineered (nano)-

particles, *Journal of Hazardous Materials*, 2020, **392**, 122335.

5 Wang, Nabi, Mohanty, Afrooz, Cantando, Aich and Baalousha, Detection and quantification of engineered particles in urban runoff, *Chemosphere*, 2020, **248**, 126070.

6 Nabi, Wang, Goharian and Baalousha, Temporal variation in TiO2 engineered particle concentrations in the Broad River during dry and wet weathers, *Science of The Total Environment*, 2021, 151081.

7 Wang, Nabi, Erfani, Goharian and Baalousha, Identification and Quantification of Anthropogenic Nanomaterials in Urban Rain and Runoff Using Single Particle-Inductively Coupled Plasma-Time of Flight-Mass Spectrometry, *Environmental Science: Nano*.

8 Nabi, Wang and Baalousha, Detection and quantification of anthropogenic titanium, cerium, and lanthanum-bearing particles home dust, *Environmental Science: Nano*.

9 Vidmar, Zuliani, Mila\vci\vc and \vS\vcan\vcar, Following the occurrence and origin of titanium dioxide nanoparticles in the Sava River by single particle ICP-MS, *Water*, 2022, **14**, 959.

10 Phalyvong, Sivry, Pauwels, Gélabert, Tharaud, Wille, Bourrat and Benedetti, Occurrence and origins of cerium dioxide and titanium dioxide nanoparticles in the Loire river (France) by single particle ICP-MS and FEG-SEM imaging, *Frontiers in Environmental Science*, 2020, **8**, 141.

11 Phalyvong, Sivry, Pauwels, Gélabert, Tharaud, Wille, Bourrat, Ranville and Benedetti, Assessing CeO2 and TiO2 nanoparticle concentrations in the Seine River and its tributaries near Paris, *Frontiers in Environmental Science*, 2021, 271.

12 Wang, Alasonati, Fisicaro and Benedetti, Titanium nanoparticles fate in small-sized watersheds under different land-uses, *Journal of Hazardous Materials*, 2022, **422**, 126695.

13 Slomberg, Auffan, Guéniche, Angeletti, Campos, Borschneck, Aguerre-Chariol and Rose, Anthropogenic release and distribution of titanium dioxide particles in a river downstream of a nanomaterial manufacturer industrial site, *Frontiers in Environmental Science*, 2020, **8**, 76.

14 Zheng, Li, Bai, Li, Hao, Yu, Cai and Liu, Spatial distribution, temporal variations and source of titanium dioxide nanoparticles in Taihu Lake, China, *Science of The Total Environment*, 2024, 175481.

15 Gondikas, Kammer, Reed, Wagner, Ranville and Hofmann, Release of TiO2 nanoparticles from sunscreens into surface waters: a one-year survey at the old Danube recreational Lake, *Environmental Science & Technology*, 2014, **48**, 5415–5422.

16 Sharma, Aggregation and toxicity of titanium dioxide nanoparticles in aquatic environment—A Review, *Journal of Environmental Science and Health, Part A*, 2009, **44**, 1485–1495.

17 Tan, Peralta-Videa and Gardea-Torresdey, Interaction of titanium dioxide nanoparticles with soil components and plants: current knowledge and future research needs-a critical

review, *Environmental Science: Nano*, 2018, **5**, 257–278.

18 Liu and Cohen, Multimedia environmental distribution of engineered nanomaterials, *Environmental Science & Technology*, 2014, **48**, 3281–3292.

19 Hochella, Mogk, Ranville, Allen, Luther, Marr, McGrail, Murayama, Qafoku, Rosso and others, Natural, incidental, and engineered nanomaterials and their impacts on the Earth system, *Science*.

20 Petosa, Jaisi, Quevedo, Elimelech and Tufenkji, Aggregation and deposition of engineered nanomaterials in aquatic environments: Role of physicochemical interactions, *Environmental Science & Technology*, 2010, **44**, 6532–6549.

21 Louie, Tilton and Lowry, Critical review: impacts of macromolecular coatings on critical physicochemical processes controlling environmental fate of nanomaterials, *Environmental Science: Nano*, 2016, **3**, 283–310.

22 Praetorius, Badetti, Brunelli, Clavier, Gallego-Urrea, Gondikas, Hassellöv, Hofmann, Mackevica, Marcomini and others, Strategies for determining heteroaggregation attachment efficiencies of engineered nanoparticles in aquatic environments, *Environmental Science: Nano*.

23 Hotze, Phenrat and Lowry, Nanoparticle aggregation: Challenges to understanding transport and reactivity in the environment, *Journal of Environmental Quality*, 2010, **39**, 1909–1924.

24 Lowry, Hill, Harper, Rawle, Hendren, Klaessig, Nobbmann, Sayre and Rumble, Guidance to improve the scientific value of zeta-potential measurements in nanoEHS, *Environmental Science: Nano*, DOI:10.1039/c6en00136j.

25 Kosmulski, Isoelectric points and points of zero charge of metal (hydr) oxides: 50 years after Parks' review, *Advances in Colloid and Interface Science*, 2016, **238**, 1–61.

26 Auffan, Pedeutour, Rose, Masion, Ziarelli, Borschneck, Chaneac, Botta, Chaurand, Labille and others, Structural degradation at the surface of a TiO2-based nanomaterial used in cosmetics, *Environmental Science & Technology*, 2010, **44**, 2689–2694.

27 Philippe, Ko\vsìk, Welle, Guigner, Clemens and Schaumann, Extraction and characterization methods for titanium dioxide nanoparticles from commercialized sunscreens, *Environmental Science: Nano*, 2018, **5**, 191–202.

28 Karlsson, Abbas, Bordes, Cao, Larsson, Taylor and Steenari, Characterisation of silicon, zirconium and aluminium coated titanium dioxide pigments recovered from paint waste, *Dyes and pigments*, 2019, **162**, 145–152.

29 Philippe and Schaumann, Interactions of dissolved organic matter with natural and engineered inorganic colloids: a review, *Environmental Science & Technology*, 2014, **48**, 8946–8962.

30 Wheeler, Chetwynd, Fahy, Hong, Tochihuitl, Foster and Lynch, Environmental dimensions of the protein corona, *Nature Nanotechnology*, 2021, **16**, 617–629.

31 Khomami, Philippe, Lechtenfeld, Guigner, Heissler, Schaumann and others, Validation of a field deployable reactor for in situ formation of NOM-engineered nanoparticle corona, *Environmental Science: Nano*, 2020, **7**, 486–500.

32 Khomami, Patel, Jusi, Trouillet, David, Schaumann and Philippe, Influential parameters of surface waters on the formation of coating onto TiO 2 nanoparticles under natural conditions, *Environmental Science: Nano*, DOI:10.1039/d1en00431j.

33 Ohtani, Prieto-Mahaney, Li and Abe, What is Degussa (Evonik) P25? Crystalline composition analysis, reconstruction from isolated pure particles and photocatalytic activity test, *Journal of Photochemistry and Photobiology A: Chemistry*, 2010, **216**, 179–182.

34 Rasmussen, Mast, De Temmerman, Verleysen, Waegeneers, Van Steen, Pizzolon, De Temmerman, Van Doren, Jensen and others, Titanium dioxide, NM-100, NM-101, NM-102, NM-103, NM-104, NM-105: characterisation and physico-chemical properties, *JRC Science and Policy Reports*.

35 Philippe, Ndoli-Kessie, Fricke, Guigner, Heider and Di Lodovico, Extraction and Characterization of TiO2 Pigments from Commercial Paints for Environmental Studies, *Separations*, 2025, **12**, 91.

36 Geochemical Atlas of Europe, http://weppi.gtk.fi/publ/foregsatlas/, (accessed 2020).

37 Gad, Khomami, Krieg, Schor, Philippe and Lechtenfeld, Environmental drivers of dissolved organic matter composition across central European aquatic systems: A novel correlation-based machine learning and FT-ICR MS approach, *Water Research*, 2025, **273**, 123018.

38 Zacharioudaki, Fitilis and Kotti, Review of fluorescence spectroscopy in environmental quality applications, *Molecules*, 2022, **27**, 4801.

39 Gabor, Baker, McKnight, Miller and others, Fluorescence indices and their interpretation, *Aquatic organic matter fluorescence*.

40 Kaufman and Rousseeuw, *Finding groups in data: an introduction to cluster analysis*, John Wiley & Sons, 2009.

41 Scofield, Hartree-Slater subshell photoionization cross-sections at 1254 and 1487 eV, *Journal of electron spectroscopy and related phenomena*, 1976, **8**, 129–137.

42 Hastie, Tibshirani and Friedman, *The elements of statistical learning: data mining, inference, and prediction*, Springer Science & Business Media, 2009.

43 Chen and Guestrin, in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

44 Lundberg, Erion, Chen, DeGrave, Prutkin, Nair, Katz, Himmelfarb, Bansal and Lee, From local explanations to global understanding with explainable AI for trees, *Nature machine*

intelligence, 2020, **2**, 56–67.

45 Goldstein, Kapelner, Bleich and Pitkin, Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation, *journal of Computational and Graphical Statistics*, 2015, **24**, 44–65.

46 Loosli, Le Coustumer and Stoll, TiO2 nanoparticles aggregation and disaggregation in presence of alginate and Suwannee River humic acids. pH and concentration effects on nanoparticle stability, *Water Research*, 2013, **47**, 6052–6063.

47 Ottofuelling, Von Der Kammer and Hofmann, Commercial titanium dioxide nanoparticles in both natural and synthetic water: comprehensive multidimensional testing and prediction of aggregation behavior, *Environmental Science & Technology*, 2011, **45**, 10045–10052.

48 Labille, Feng, Botta, Borschneck, Sammut, Cabie, Auffan, Rose and Bottero, Aging of TiO2 nanocomposites used in sunscreen. Dispersion and fate of the degradation products in aqueous environment, *Environmental Pollution*, 2010, **158**, 3482–3489.

49 Pallem, Stretz and Wells, Evaluating aggregation of gold nanoparticles and humic substances using fluorescence spectroscopy, *Environmental Science & Technology*, 2009, **43**, 7531–7535.

50 Hunter, *Foundations of Colloid Science* , Oxford University Press, Second Edition., 2001.

51 Verrall, Warwick and Fairhurst, Application of the Schulze-Hardy rule to haematite and haematite/humate colloid stability, *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, 1999, **150**, 261–273.

52 Beckett and Le, The role or organic matter and ionic composition in determining the surface charge of suspended particles in natural waters, *Colloids and Surfaces*, 1990, **44**, 35–49.

53 Beauvois, Vantelon, Jestin, Rivard, Bouhnik-Le Coz, Dupont, Briois, Bizien, Sorrentino, Wu and others, How does calcium drive the structural organization of iron-organic matter aggregates? A multiscale investigation, *Environmental Science: Nano*, 2020, **7**, 2833–2849.

54 Liu, Wazne, Chou, Xiao and Xu, Influence of Ca2+ and Suwannee River Humic Acid on aggregation of silicon nanoparticles in aqueous media, *Water Research*, 2011, **45**, 105–112.

55 Walch, Praetorius, von der Kammer and Hofmann, Generation of reproducible model freshwater particulate matter analogues to study the interaction with particulate contaminants, *Water Research*, 2023, **229**, 119385.

56 Liu, Legros, von der Kammer and Hofmann, Natural Organic Matter Concentration and Hydrochemistry Influence Aggregation Kinetics of Functionalized Engineered Nanoparticles, *Environmental Science & Technology*, 2013, **47**, 4113–4120.

57 Liyuan, 1988.

58 Heil and Sposito, Organic matter role in illitic soil colloids flocculation: II. Surface charge, *Soil Science Society of America Journal*, 1993, **57**, 1246–1253.

59 Liu, Sun, Borthwick and Ni, Comparison on aggregation and sedimentation of titanium dioxide, titanate nanotubes and titanate nanotubes-TiO2: Influence of pH, ionic strength and natural organic matter, *Colloids and Surfaces A: Physicochemical and Engineering Aspects*, 2013, **434**, 319–328.

60 Zhang, Chen, Westerhoff and Crittenden, Impact of natural organic matter and divalent cations on the stability of aqueous nanoparticles, *Water Research*, 2009, **43**, 4249–4257.

61 Nickel, Hellack, Nogowski, Babick, Stintz, Maes, Schäffer and Kuhlbusch, Mobility, fate and behavior of TiO2 nanomaterials in different environmental media, *Environmental Research of the Federal Ministry for the Environment*.

62 Li, Yang, Guo, Dang, Li and Zhang, Effects of humic acids on the aggregation and sorption of nano-TiO 2, *Chemosphere*, 2015, **119**, 171–176.

63 Degenkolb, Kaupenjohann and Klitzke, The Variable Fate of Ag and TiO2 Nanoparticles in Natural Soil Solutions—Sorption of Organic Matter and Nanoparticle Stability, *Water, Air, and Soil Pollution*, 2019, **230**, 62.

64 Yan, Sharma, Chen, Li and Shang, Coupled impact of proteins with different molecular weights and surface charges on TiO 2 mobility, *Environmental Science: Nano*, 2022, **9**, 2773–2787.

65 Shakiba, Hakimian, Barco and Louie, Dynamic Intermolecular Interactions Control Adsorption from Mixtures of Natural Organic Matter and Protein onto Titanium Dioxide Nanoparticles, *Environmental Science & Technology*, 2018, **52**, 14158–14168.

66 Parks, The isoelectric points of solid oxides, solid hydroxides, and aqueous hydroxo complex systems, *Chemical Reviews*, 1965, **65**, 177–198.

67 Thio, Zhou and Keller, Influence of natural organic matter on the aggregation and deposition of titanium dioxide nanoparticles, *Journal of Hazardous Materials*, 2011, **189**, 556–563.

68 Li and Sun, A comparative study on aggregation/sedimentation of TiO2 nanoparticles in mono-and binary systems of fulvic acids and Fe (III), *Journal of Hazardous Materials*, 2011, **197**, 70–79.

Data availability statement

The raw-data for this article as well as the complete R-code used for the data analysis, the modeling, and the sites selection are available at Zenodo at https://zenodo.org/records/14965814.

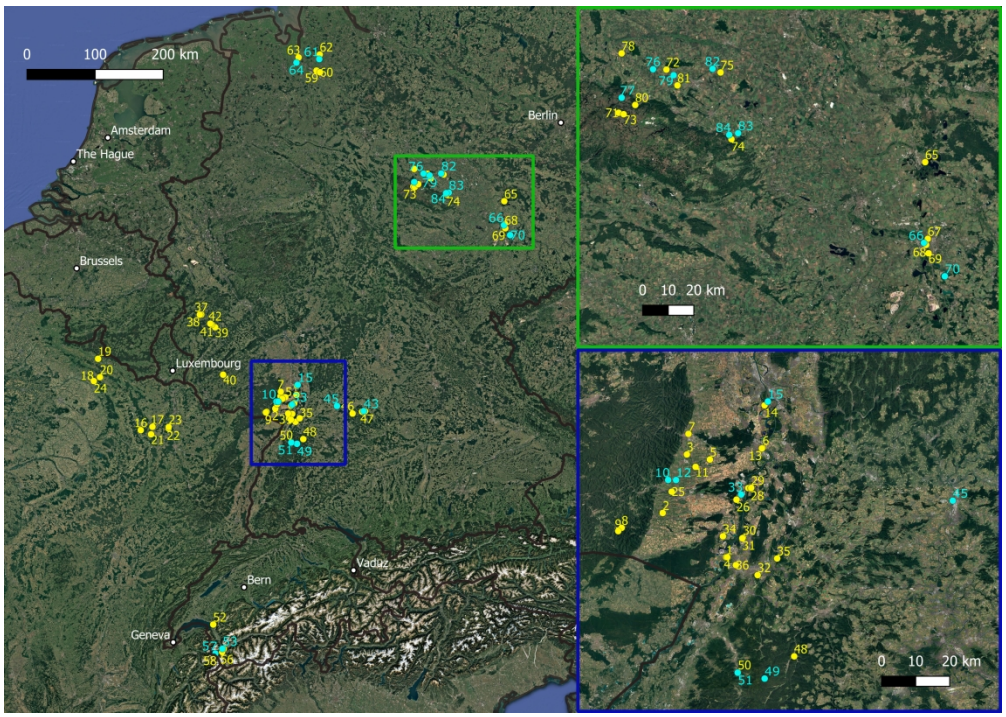


Figure 1: map of the sites screened in this study (yellow dots) and the sites selected for the exposure experiments (cyan dots). Each site was identified with a number. The map was generated using QGIS with Google satellite images.

296x209mm (300 x 300 DPI)

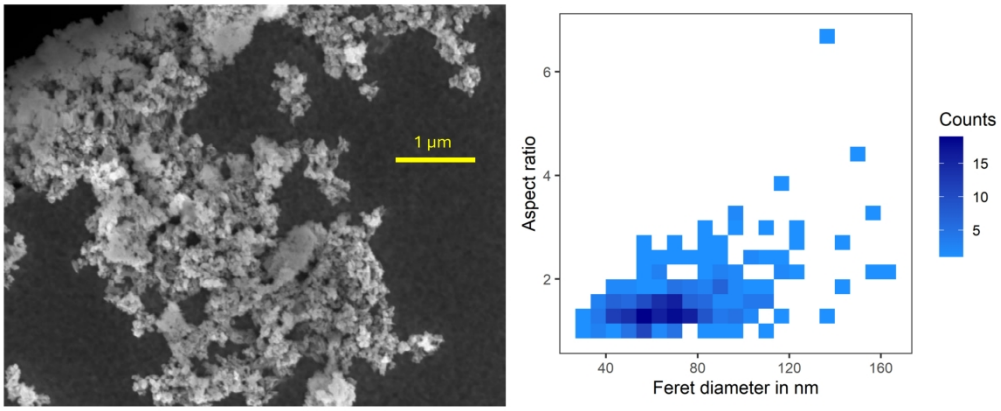


Figure 2: representative SEM image (left) and size and aspect ratio distribution (n = 306) of the TiO_2-nanoparticles extracted from the commercial sunscreen and used in this study (right).

316x130mm (300 x 300 DPI)

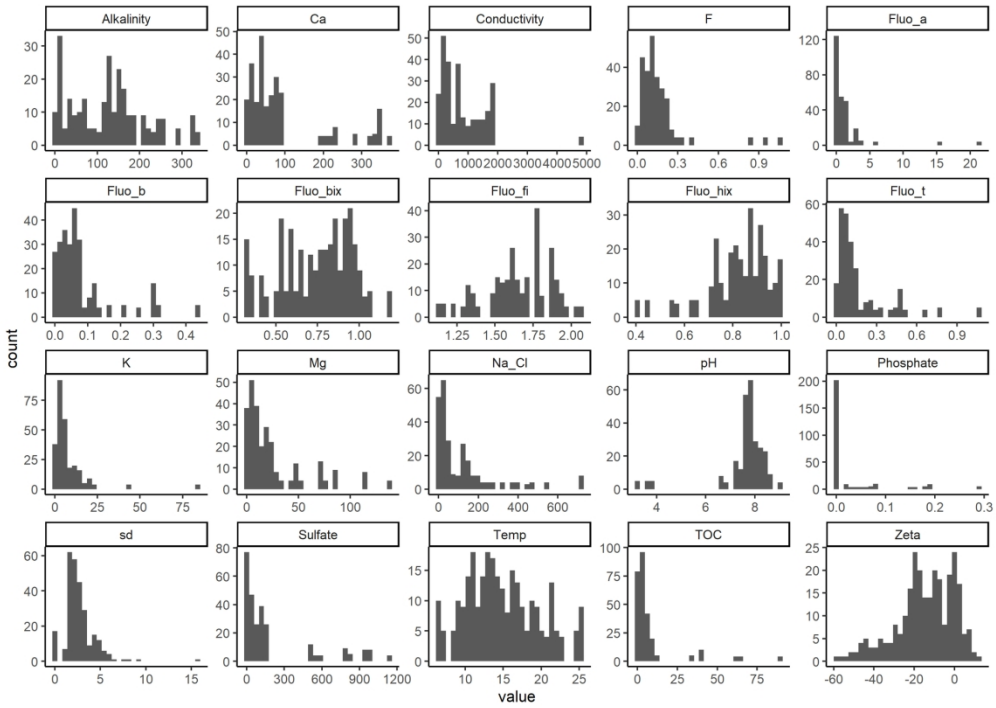


Figure 3: density distribution (n = 269) of all numerical parameters in the dataset used in this study. The parameters are sorted in alphabetical order. Ca, F, K, Mg, Phosphate, and Sulfate denote the concentration of the corresponding ions in mg/L. Na_Cl is the sum of Na^+ and Cl^- concentrations in mg/L. Conductivity is given in µS cm-1, total dissolved organic carbon (TOC) in mg/L, ζ-potential and the standard deviations (sd) of the individual ζ-potential measurements (n = 10) in mV, and average water temperature during the exposure in °C. The parameters named Fluo_X are unitless descriptors of the organic matter composition obtained from the fluorescence maps as explained in the method description.

177x127mm (300 x 300 DPI)

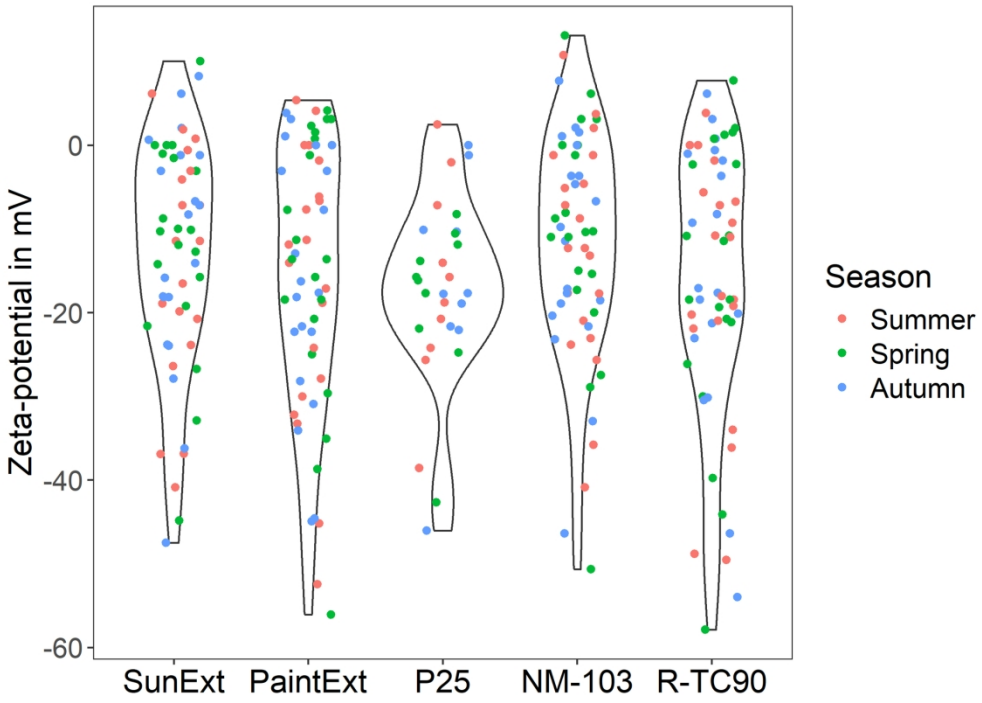


Figure 4: distribution of measured ζ-potential values (n = 269) as a function of particles type (x-axis) and Season (color).

177x127mm (300 x 300 DPI)

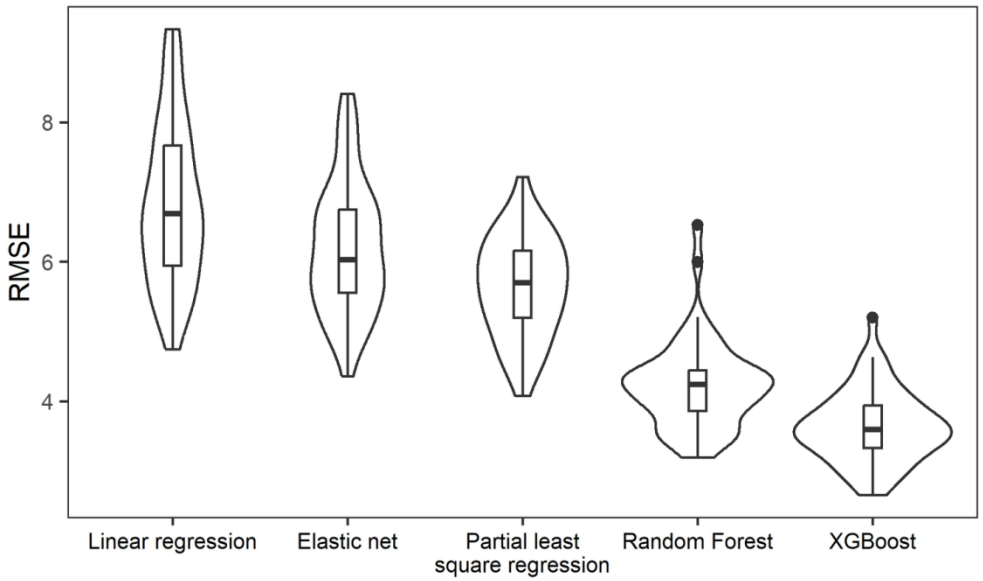


Figure 5: model performances comparison (root mean square error on the test dataset, RMSE) of the models tested to predict ζ-potential. These values are much lower than the standard deviation of the ζ-potentials in the whole dataset was 14.5 mV indicating the mean error corresponding to a random guessing process. The distributions represent the RMSE of 50 models generated from different random training/test data-partitions.

127x76mm (300 x 300 DPI)

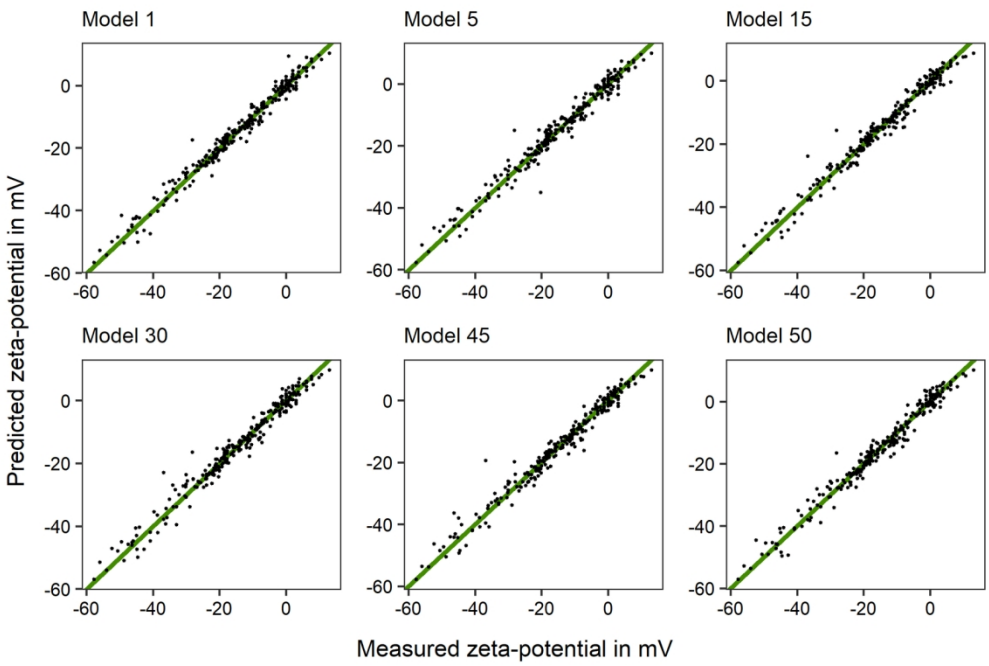


Figure 6: ζ-potentials values predicted by the XGBoost model versus the measured values for a selection of six among 50 randomly generated models. The green line denotes the y = x function. All datapoints were used for these plots. The different models are consistent in terms of accuracy. The very good alignment of the data points with the y = x lines illustrates the excellent prediction power as indicated by the low RMSEs.

152x101mm (300 x 300 DPI)

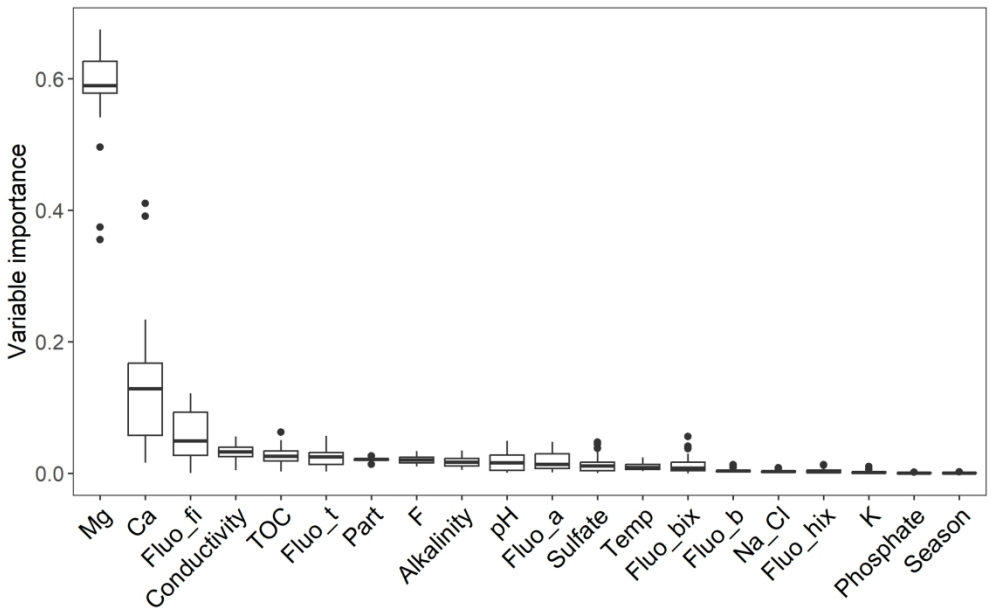


Figure 7: variables importance of the input parameters (water composition, particle type, temperature, and season) for the accuracy of the prediction of ζ-potentials by the 50 randomly generated XGBoost models. Ca, F, K, Mg, Phosphate, and Sulfate: concentration of the corresponding ions. Na_Cl: sum of Na^+ and Cl^- concentrations. TOC: total dissolved organic carbon (TOC), Temp: average water temperature during the exposure. The parameters named Fluo_X are descriptors of the organic matter composition obtained from the fluorescence maps as explained in the method description. Part: particle type.

203x127mm (300 x 300 DPI)

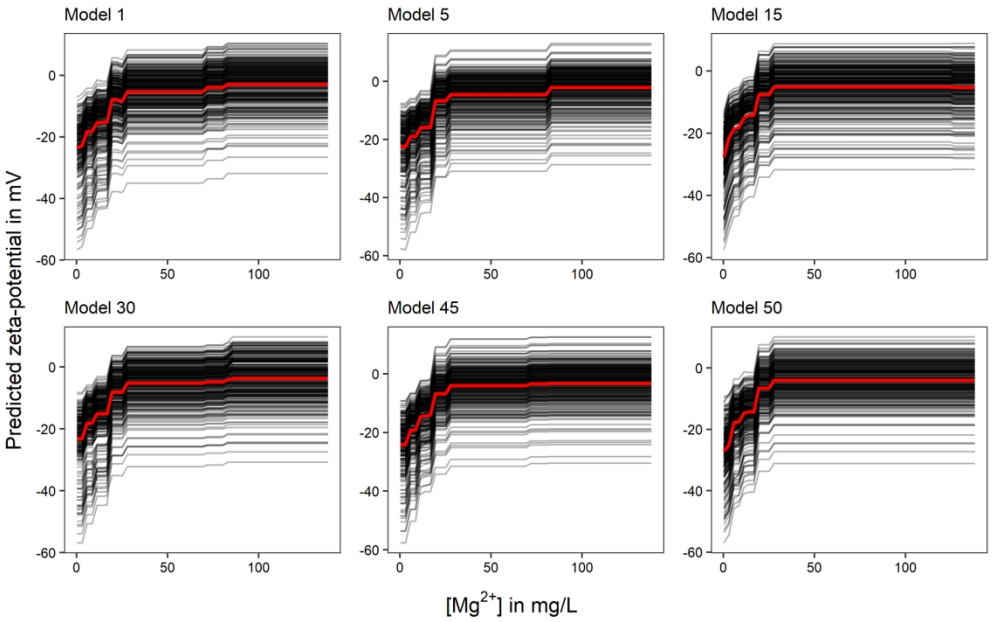


Figure 8: partial dependence (red lines) and individual conditional expectation (grey lines) plots of the input variable [Mg^{2+}] for a selection of six among 50 randomly generated XGBoost models.

203x127mm (300 x 300 DPI)

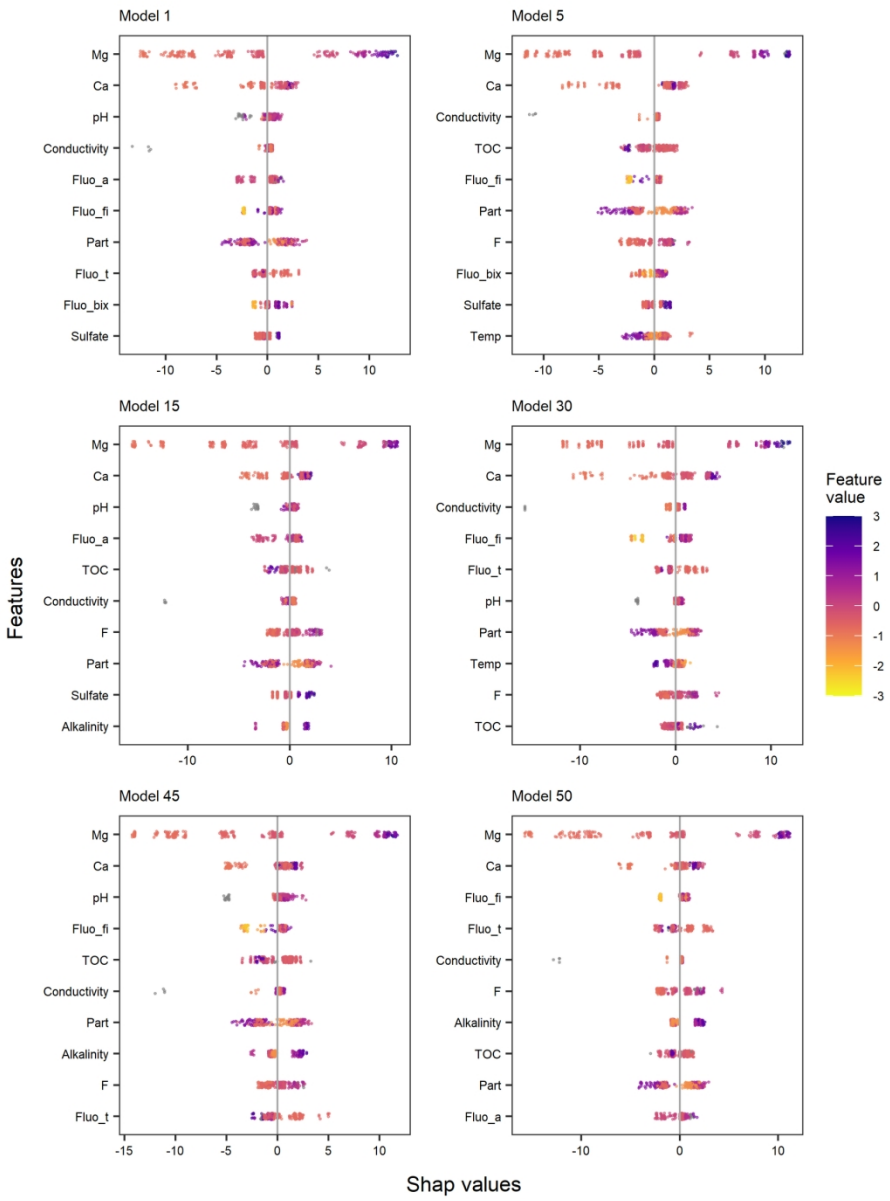


Figure 9: Shapley (Shap) values distributions (x-axis) for the ten most important variables (y-axis) for a selection of six among 50 randomly generated XGBoost models. The Shap values are the natural logarithm of the odds shift of the prediction due to the evaluated variable. A high negative or positive value denotes a high probability that the predicted output value will be low or high because of the input variable. The input variables are sorted by importance. The color legend denotes the normalized values of the input variables, the darker the higher the value. For the particle type ("Part"), a numerical factor from 1 to 5 was assigned to the five types of particles in the following order: sunscreen extract, paint extract, P25, NM-103, and R-TC90.

152x203mm (300 x 300 DPI)

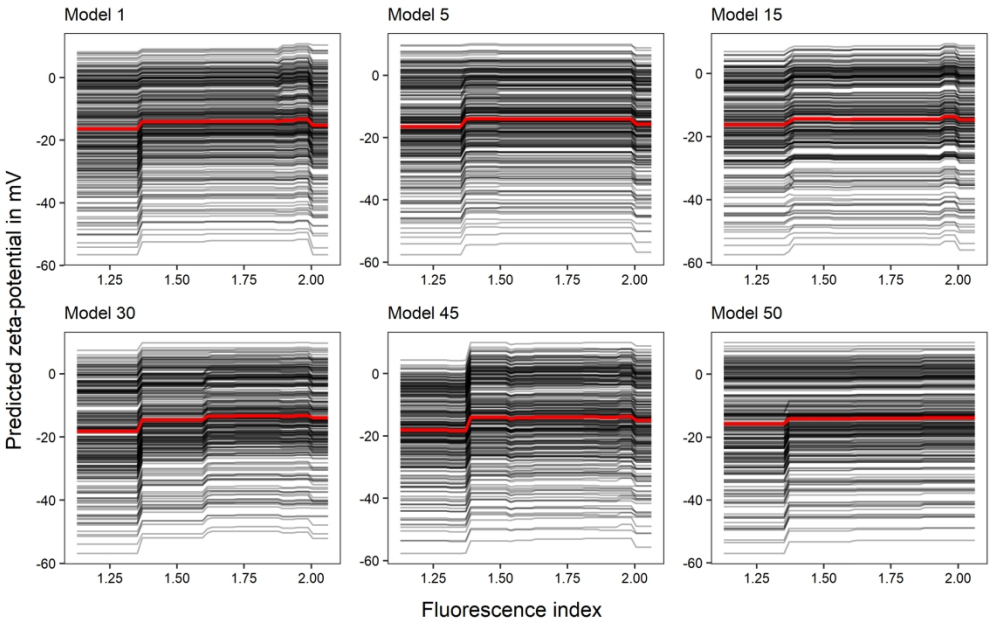


Figure 10: partial dependence (red lines) and individual conditional expectation (grey lines) plots of the fluorescence index for a selection of six among 50 randomly generated XGBoost models.

203x127mm (300 x 300 DPI)

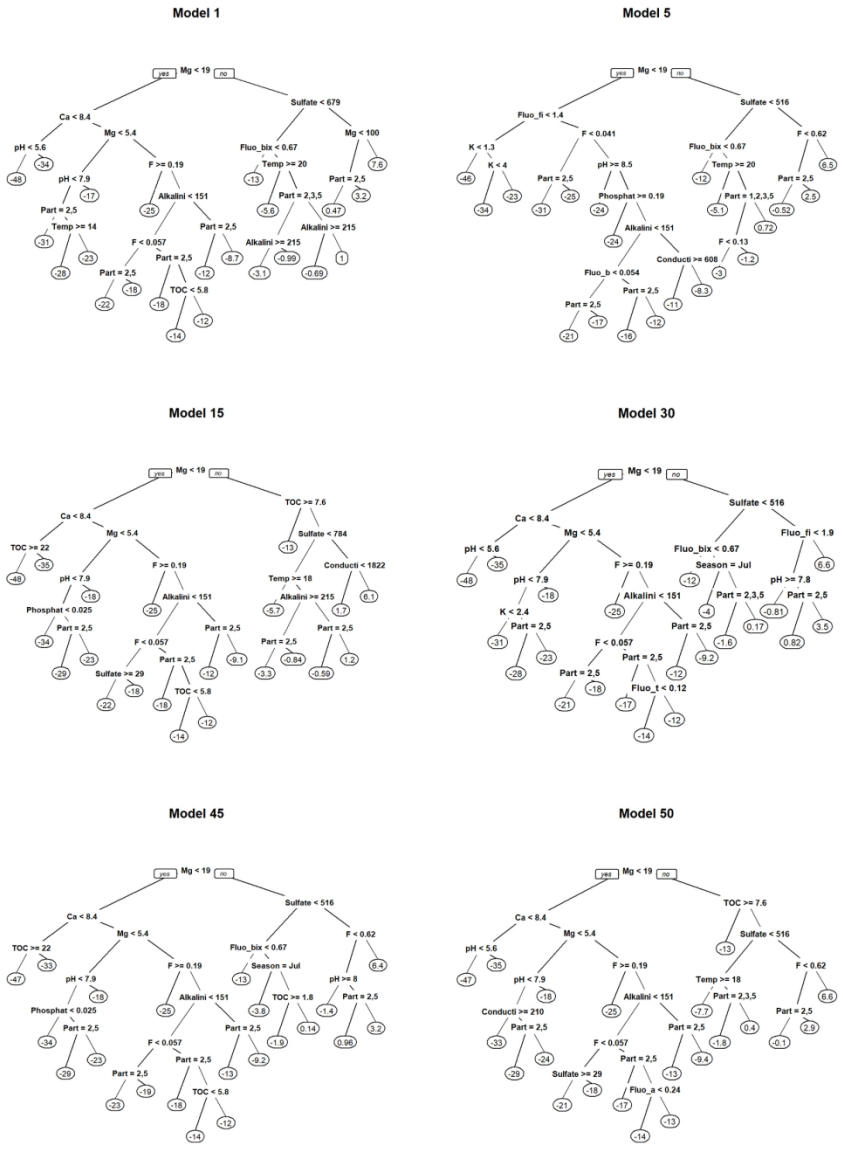


Figure 11: surrogate decision trees generated for a selection of six among 50 randomly generated XGBoost models. The leaves of the trees are the predicted ζ-potentials and the numbers at the nodes next to the input variables' names denote the threshold values for the split decision. For the categorical variable particle type ("Part"), a numerical factor from 1 to 5 was assigned to the five types of particles in the following order: sunscreen extract, paint extract, P25, NM-103, and R-TC90.

152x203mm (300 x 300 DPI)

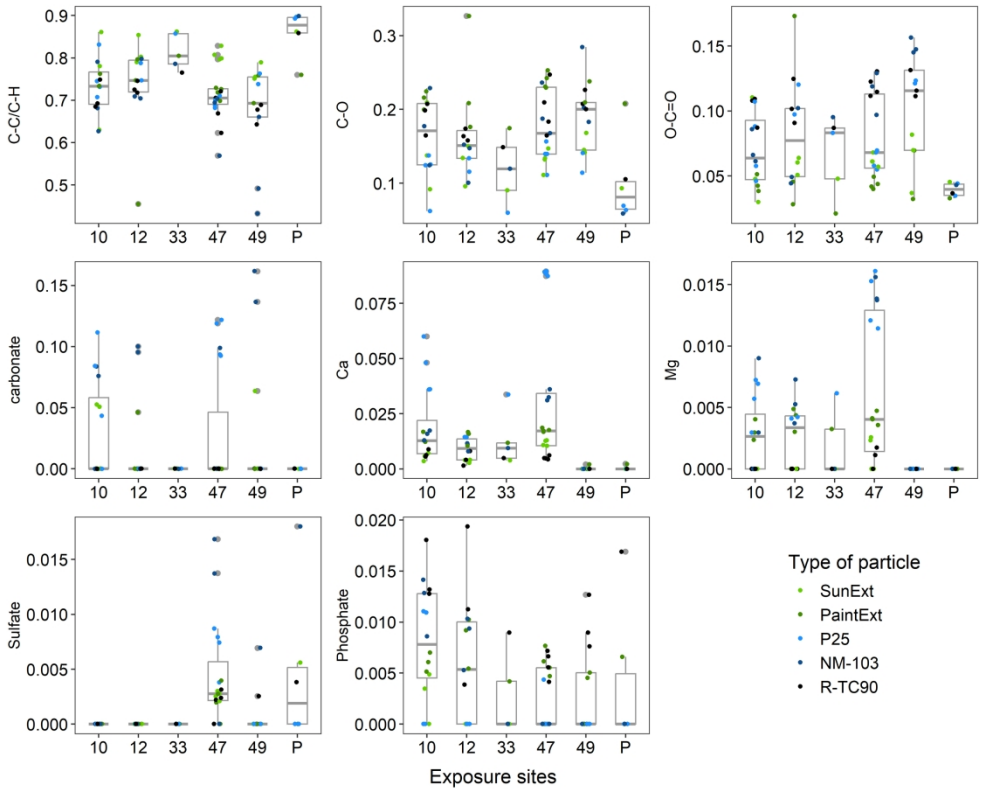
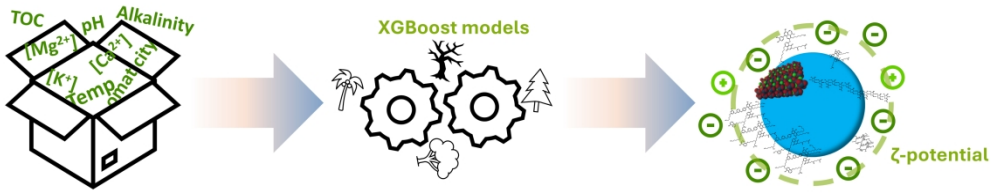


Figure 12: fractional contribution of each element detected on the TiO_2-nanoparticles before (P) and after exposure (numbers) to a selection of contrasting surface waters. The ratios are calculated based on the XPS quantitative analysis. For better comparability, all concentrations (at%) are normalized by the total carbon content concentrations of the elements in atomic %. See text for the surface waters' description.

254x203mm (300 x 300 DPI)



316x60mm (300 x 300 DPI)