Environmental Science
Water Research & Technology

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal

Cite this: DOI: 10.1039/d5ew01176k

Predictive models for the formation of emerging disinfection by-products

Argyri Kozari * and Voutsa Dimitra*

Disinfection is a necessary process during drinking water treatment; however, there is potential risks due to the formation of disinfection by-products (DBPs). Besides the regulated DBPs (THMs and HAAs), there is a growing concern about emerging DBPs (eDBPs), which appear to be much more toxic than the regulated ones. Climate-driven conditions can affect the concentrations and profiles of DBPs and enhance the formation of eDBPs. Since eDBPs are not included in the current monitoring programs, there is a need to predict their concentrations in drinking water using modeling techniques. The aim of this study is the development and application of a predictive modeling framework for the formation of selected eDBPs (haloacetonitriles, haloacetamides, halonitromethanes and haloketones) under chlorination or chloramination conditions in drinking water treatment systems. Through multiple regression analysis, linear and logarithmic models are used based on a dataset that describes the formation of eDBPs under various climatic conditions (seawater intrusion, flooding events, algal blooming and enrichment by humic acid) that could impact water sources. The dataset includes information on the concentrations of eDBPs, water quality parameters and disinfection conditions. Besides DOC and UV_{254}, different forms of nitrogen appear to be a significant predicting parameters on the formation of eDBPs, as well as the dose and contact time of disinfection. The proposed relationships can be useful tools for the identification and control of eDBPs in drinking water.

Received 26th November 2025,
Accepted 13th March 2026

DOI: 10.1039/d5ew01176k

rsc.li/es-water

Water impact

The evaluation of exposure to non-monitored DBPs is crucial because of their toxicity. Predictive models can be useful tools for the management and regulation of drinking water by identifying and controlling the presence of eDBPs under a wide range of water quality variations due to climate changes, contributing to the availability of safe drinking water worldwide.

1. Introduction

Chlorination is a widely used disinfection process for drinking water treatment to eliminate pathogenic microorganisms and protect public health. However, the reactions of chlorine with organic matter, anthropogenic contaminants, inorganic bromide and iodide, and other constituents present in source water lead to the formation of disinfection by-products (DBPs). Since the discovery of THMs in the 1970s, more than 700 DBPs have been identified. However, only a small number of them are regulated in developed countries. Currently, the European Directive regulates four THMs, five haloacetic acids (HAAs), bromate and chlorite.[1] There are many factors that influence the formation of DBPs, such as disinfectant dose, contact time, concentration and characteristics of natural organic matter (NOM), and the presence of nitrogen, inorganic anions and anthropogenic contaminants.[2,3] Regulated THMs and HAAs have been extensively studied, and many predictive models have been developed.[4]

Nowadays, there is a growing concern over the formation of emerging DBPs (eDBPs), including halonitromethanes, haloacetamides, haloacetonitriles and haloketones, as they lack well-characterized formation pathways.[2,5] Many of these eDBPs exhibit higher cytotoxicity and genotoxicity than the regulated DBPs, raising significant public health concerns.[6–8]

Recent researches have focused on elucidating the formation mechanisms of nitrogenous DBPs (N-DBPs), which are typically more cytotoxic and genotoxic than their carbonaceous analogues.[8] Unregulated haloacetonitriles, particularly dihaloacetonitriles, are important toxicity drivers in drinking water from various sources across the United States.[9] Various compounds from the class of haloacetonitriles (dichloroacetonitrile and

Environmental Pollution Control Laboratory, School of Chemistry, Aristotle University, 541 24 Thessaloniki, Greece. E-mail: akozaria@chem.auth.gr, dvoutsa@chem.auth.gr

dibromoacetonitrile), haloacetamides (dichloroacetamide and dibromoacetamide), and halonitromethanes (trichloronitromethane and bromonitromethane) are included in the prioritization list of DBPs based on specific criteria, incorporating regulatory standards set by public health organizations and considering the observed and probable severity of health impacts.[10]

Moreover, the impact of climate change (*e.g.*, extensive droughts, extreme floods, increasing frequency of algal blooms, and seawater intrusion), use of reclaimed water in areas with water scarcity, formation of DBPs from organic micropollutants that are not well-removed during wastewater treatment, as well as the advanced treatment procedures, introduce even more complex reactions with chlorine, resulting in changes in patterns and formation of new classes of DBPs.[3,11,12] Since eDBPs are not included in any current monitoring program for drinking water, the information regarding their concentrations, as well as the factors affecting their formation, are limited.

A variety of modeling approaches have been developed to estimate DBP formation under different treatment conditions. Some water quality parameters, such as total organic carbon and UV absorption, can be monitored in real time. Making use of these parameters, models have been developed to predict the concentrations of DBPs in drinking water, thus enabling real-time estimation.[13–16] Empirical models, particularly multiple linear regression models, have been used to relate DBP concentrations to water quality parameters such as dissolved organic carbon (DOC), UV_{254}, and specific ultraviolet absorbance (SUVA).[17] There are, however, cases in which the models show significant deviations from actual concentrations.[17–19] Usually, these models do not take into account other parameters that can affect the formation of DBPs, such as water disinfection conditions and additional water parameters.[20] Furthermore, the site-specific variability in NOM composition and water matrix characteristics reduces the generalizability of models trained on limited datasets.[21]

The existing models focus on the prediction of regulated DBPs, *i.e.*, THMs and HAAs.[4,22–30] On the other hand, the proposed models for eDBPs are rare. The data on eDBPs are very limited, as these byproducts are not monitored regularly since there are no legislative obligations. Moreover, there is a major difficulty in predicting eDBPs, particularly under dynamic treatment scenarios and in systems with highly variable source water quality, as climate change worsens the predictability of these compounds.[31,32]

Understanding and predicting the formation of eDBPs remains a complex task due to their diverse chemical structures, variable precursor sources, and multiple influencing factors, such as disinfectant type, pH, temperature, contact time, and bromide/iodide content.[5,20] eDBPs are often formed from nitrogen-containing precursors, such as amino acids and proteins, which are more variable and less predictable under extreme weather events than the humic substances associated with regulated DBPs.[33,34] As a result, traditional empirical models based only on parameters like DOC or UV_{254} often fail to capture the complex dynamics of eDBP formation,[35] and at the same time, some existing models are based on a limited number of samples due to a lack of datasets. Developing large and comprehensive datasets for unregulated DBPs is of great importance, as improved modeling strategies can accommodate the diversity of eDBP precursors and formation conditions and adapt to challenging real-world treatment environments. Table 1 presents the existing literature on predictive models for unregulated DBPs, including the current study.

This study aims to develop and apply a predictive modeling framework for the formation of selected emerging DBPs (haloacetonitriles, haloacetamides, halonitromethane,

Table 1 Literature review on the predictive models for assessing the regulated and unregulated formation of DBPs

DBPs	No. of samples	Water source	Disinfection	Independent variables	References
HANs, HKs	N/A	DWTP	Chlorination	*T*, pH, conductivity, turbidity, DOC, UV_{254}	43
THMs, HAAs, NDMA, DCAN	425–1306	Distribution system	Chlorination	pH, alkanite, turbidity, F^-, DOC, chlorine dose	44
HKs	50	Distribution system	Chlorination	*T*, pH, UV_{254}, DOC, Br^-, NH_4^+, NO_2^-, residual free chlorine	41
THMs, HAAs, HANs	168–183	EfOM, AOM, NOM	Chlorination	pH, UV_{254}, DOC, Br^-, DON/DOC, chlorine dose	42
THMs, HAAs, HANs, HKs	N/A	Raw water, DWTP (prior and post-disinfection)	Chlorination/chloramination	*T*, pH, UV_{254}, DOC, SUVA, Br^-, contact time	21
THMs, HAAs, HANs, TCNM, HKs	>150	Distribution systems	Chlorination	*T*, pH, residual chlorine, DOC, turbidity, conductivity, UV_{254}	15
NDMA	220	DWTP	Chlorination	*T*, retention time, monochloramine residue	45
THMs, HAAs, HANs	N/A	Distribution system	Chlorination	pH, residual chlorine, NH_3, SUVA	13
HANs, TCNM, HAcAms, HKs	221	Surface water	Chlorination/Chloramination	DOC, UV_{254}, SUVA, NO_2^-, NO_3^-, NH_4^+, DON, TN, chlorine dose, contact time	Current study

N/A: not available.

Table 2 Studied regimes and experimental conditions

Climate-driven scenarios	Studied regimes	Number of samples	Description[a,b]	Chlorination/chloramination conditions	
				Dose (mg L^{-1})	Contact time (h)
River	RI	69	River water	5, 10	24, 72
Sea water intrusion	SW	64	River water + sea water (contribution 0.5, 1, 2, 4%)	5, 10	24, 72
Flooding events	RA	32	River water + rainwater (moderate/high/very high flooding)	5, 10	24, 72
Water blooming (*Anabaena*)	AN	32	River water + algal organic carbon (contribution 15% and 30% to the DOC of river water)	5, 10	24, 72
Enrichment by humic substances	HA	32	River water + humic acids (15% and 20% contribution of humic organic carbon to the DOC of river water)	5, 10	24, 72

[a] Kozari and Voutsa, 2023 (ref. 12). [b] Kozari *et al.*, 2024 (ref. 11).

and haloketones) in drinking water treatment systems based on routinely monitored water quality parameters, including various forms of nitrogen that affect the formation of eDBPs and various disinfection conditions. For this purpose, multivariate linear and logarithmic regression models were applied. By integrating all the experimental data with statistical modeling techniques, this research can aid in the enhancement of the prediction of eDBPs formation and optimization of treatment processes for improved water quality outcomes.

2. Experimental

2.1. Experimental dataset

To set up the models for the prediction of eDBPs, experimental data from our previous studies were used.[11,12] Those studies reported possible impacts of climate change on the water quality of the Aliakmon River, and consequently on the formation of eDBPs. Aliakmon River is the main source of drinking water and supplies, after treatment, with 150 000 m^3 water per day, to the city of Thessaloniki, Greece. Four different cases of climatic changes that could affect the river water characteristics, namely, sea level rise, flooding events, water blooming and the enrichment of source water by humic substances, were investigated under laboratory conditions for the predictive model. Table 2 summarizes all the information about the studied regimes. Under these cases, water has been chlorinated or chloraminated with different doses and different contact times, according to the simulated distribution system test that simulates the formation of disinfection by-products under conditions that occur in real water distribution systems.[36,37] The disinfection experiments were studied at two different contact times (24 h and 72 h) with two chlorine doses of 5 and 10 mg L^{-1}. The experiments were conducted at a temperature of 20.0 ± 1.0 °C and pH 7.8 ± 0.2; the samples were buffered with 10 mM phosphate buffer and kept free of headspace. NH_4Cl was used as a quenching agent to end the chlorination process, and ascorbic acid was used for the termination of chloramination. The analysis of DBPs was based on modified U.S. EPA Method 551.1. Briefly, DBPs were extracted by liquid–liquid extraction with MTBE, and the extracts were analyzed by gas chromatography equipped with an electron capture detector (GC/ECD, Trace GC Ultra, Thermo Scientific).[36,37] The target eDBPs were 4 haloacetonitriles-HANs (TCAN, DCAN, BCAN, and DBAN), 1 halonitromethane-HNM (TCNM), 3 haloacetamides-HAcAms (CAcAm, DCAcAm, and BAcAm) and 2 haloketones-HKs (DCP and TCP).

Water quality parameters were also determined. These include dissolved organic carbon DOC, absorbance UV_{254}, specific ultraviolet absorbance SUVA, nitrite NO_2^-, nitrate NO_3^-, ammonium NH_4^+, total organic nitrogen TON and total nitrogen TN.

The experimental data represent a wide range of water quality parameters (DOC, forms of nitrogen, *etc.*) even in cases of extreme conditions, including the levels of eDBPs and disinfection conditions, and were used as a database (n = 221 samples) for the development of predictive models. The range of water parameters and eDBPs during the chlor(am)ination process is presented in Fig. 1.

2.2. Multivariate predictive model

Two multivariate predictive models were employed, one linear and one logarithmic (ln-linear), derived from the regression analysis for each compound and class of compounds. All data from the experimental procedures, carried out on a laboratory scale, were used to create the models. For the statistical analysis, Microsoft Excel and IBM SPSS Statistics software were used. The correlation analysis was based on Spearman's rho. The following water parameters (and their respective logarithmic values) were considered as independent variables in the model design: DOC, UV_{254}, SUVA, NO_2^-, NO_3^-, NH_4^+, DON, TN, as well as the disinfection conditions: disinfectant dose (dose) and disinfection contact time (CT). The reason for selecting a wide range of independent parameters is to represent possible impacts of climate change on water

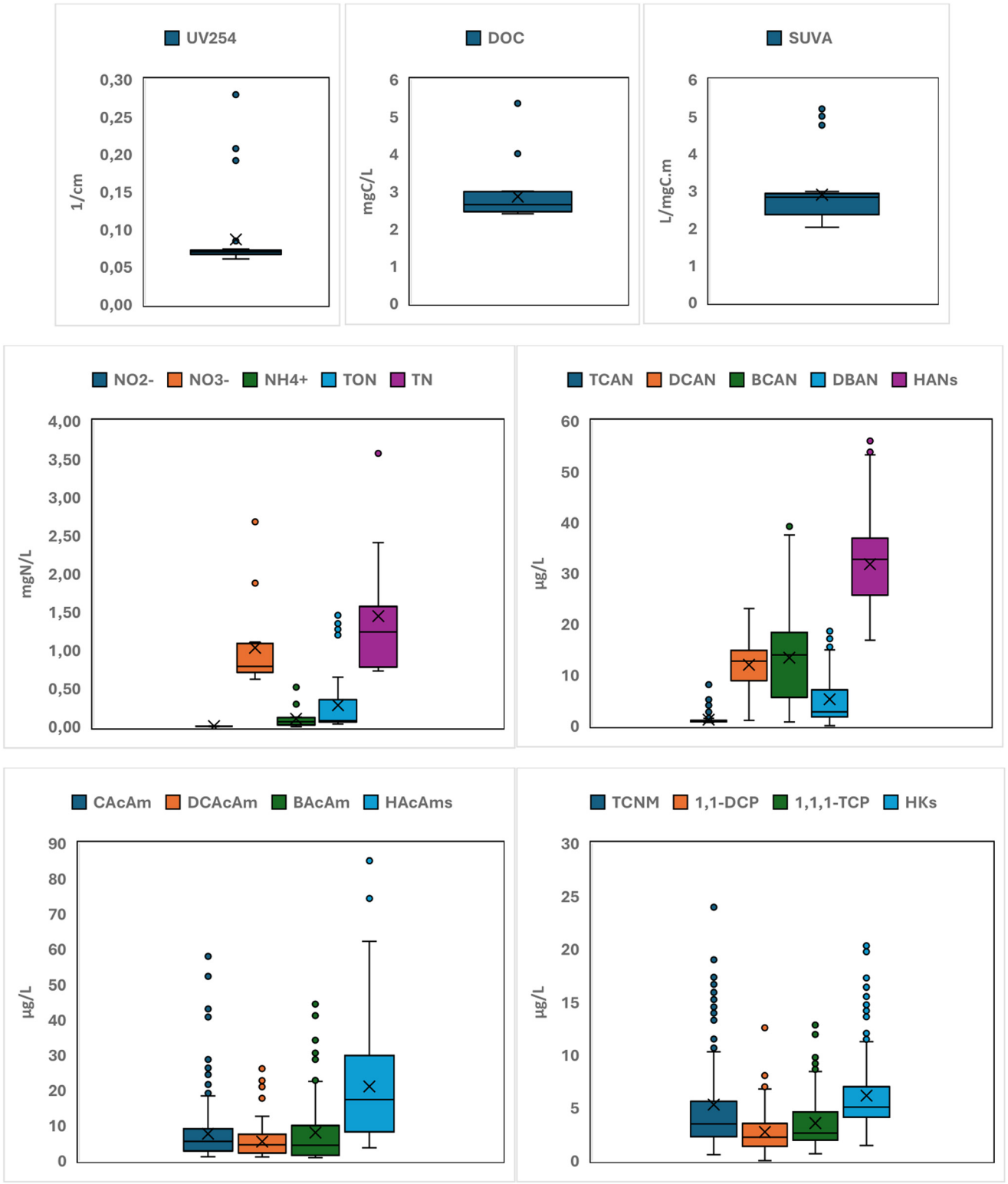

Fig. 1 Box and whisker plots for each variable in the datasets. Bottom and top of the boxes represent the 25th and 75th percentiles, respectively. The bottom and top of the whiskers represent the 10th and 90th percentiles, respectively. Lines inside the boxes represent median values, the "×" symbol represents mean values, and dots represent outliers.

quality. Climate change affects not only the profiles of carbon and nitrogen species in source water but also contributes to bacterial growth and resistance to disinfectants.[38–40] The dependent variables were the concentrations of DBPs and their respective logarithmic values (for the ln-linear model). The establishment of

relationships was based on multiple regression analysis, and the investigation of correlations and regression coefficients was performed using Anova.

For the development of the models, the correlations between the eDBPs (dependent variables) and the independent variables were investigated step-by-step. In every step, the parameters that presented statistically insignificant correlations were excluded. Only when a parameter statistically significant correlated with the dependent variable, it was considered as an independent variable for the model. The regression coefficients for each independent variable and the final relationship between the dependent variable and the independent variables were obtained by Anova. Fig. 2 and 6 present the significance of correlation between independent and dependent variables in the form of heatmaps.

The prediction performance of the proposed relationships, as relative error (%), was evaluated based on the percentage difference between the actual measurements and the predicted values.

$$\text{Rel. Error} = \frac{\text{MV} - \text{PV}}{\text{MV}} \times 100$$

MV is the concentration determined experimentally, and PV is the predicted concentration.

At the same time, the Bland–Altman plot was employed to compare the experimentally determined concentrations with the calculated values based on the proposed models. The Bland–Altman plot is a useful tool for analytical chemistry, environmental science, biomedicine, *etc.*, as it is a method of data plotting used in analyzing the agreement between two different datasets.[46,47] It plots the difference between the experimental data and data calculated from the models on the *Y*-axis against their average value on the *X*-axis. The upper and lower limits, obtained as ±1.96 standard deviations of the differences from the mean difference, are used to identify outliers in the data. A good agreement is considered when approximately 95% of data points fall between these limits.

Model generalizability was assessed by applying both internal and external model validation. Additionally, the non-parametric statistical Wilcoxon rank test was performed to compare the dataset of experimental concentrations of DBPs with the dataset of calculated concentrations from predictive models.

3. Results and discussion

3.1. Linear models

The linear model is based on the following relationship:

$$y = \text{constant} + a \times d + b \times e + c \times f + \ldots,$$

where y is the concentration of the individual eDBP or the eDBP class; d, e, and f are independent variables (concentrations of the physico-chemical parameters of water and the disinfection conditions) and a, b, and c are the regression coefficients.

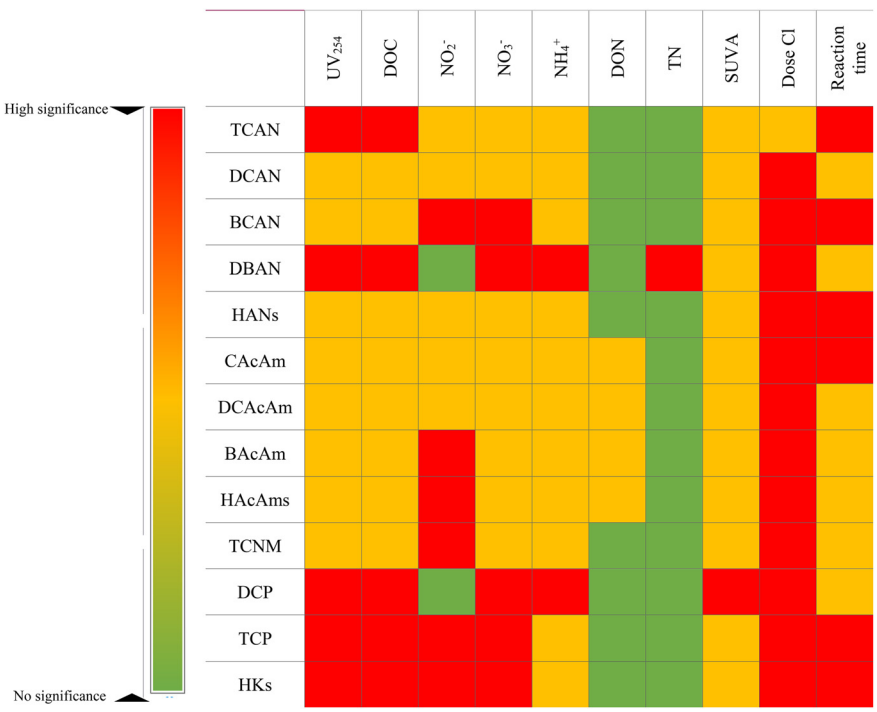

Fig. 2 Significance of the correlation between independent and dependent variables in linear models. The parameters that presented no statistically significant correlation were excluded during the development of models.

The resulting linear formation prediction models for each compound and class of eDBPs are

$$\mathbf{TCAN} = 16.139 - 178.933 \times UV_{254} - 1.325 \times DOC + 77.645 \times NO_2^- + 0.221 \times NO_3^- - 1.298 \times NH_4^+ - 0.091 \times SUVA + 0.063 \times Dose - 0.001 \times CT$$

$$\mathbf{DCAN} = 102.31 - 1202.572 \times UV_{254} + 19.231 \times DOC - 788.626 \times NO_2^- - 42.599 \times NO_3^- - 31.820 \times NH_4^+ - 7.047 \times SUVA + 0.341 \times Dose - 0.009 \times CT$$

$$\mathbf{BCAN} = -118.983 + 1461.218 \times UV_{254} + 42.614 \times DOC - 6002.724 \times NO_2^- - 84.767 \times NO_3^- + 57.095 \times NH_4^+ + 3.838 \times SUVA + 0.785 \times Dose + 0.099 \times CT$$

$$\mathbf{DBAN} = 287.657 - 1530.130 \times UV_{254} - 112.364 \times DOC + 136.158 \times NO_3^- - 131.057 \times NH_4^+ + 17.310 \times TN - 0.933 \times SUVA + 0.396 \times Dose - 0.006 \times CT$$

$$\mathbf{HANs} = 221.322 - 1951.795 \times UV_{254} - 18.838 \times DOC - 727.295 \times NO_2^- - 2.576 \times NO_3^- - 3.483 \times NH_4^+ - 4.995 \times SUVA + 1.562 \times Dose + 0.082 \times CT$$

$$\mathbf{CAcAm} = 39.700 - 597.699 \times UV_{254} - 0.116 \times DON + 649.874 \times NO_2^- + 1.891 \times NO_3^- + 19.286 \times NH_4^+ - 1.931 \times SUVA + 0.464 \times Dose + 0.061 \times CT$$

$$\mathbf{DCAcAm} = -63.935 + 455.916 \times UV_{254} + 8.401 \times DON + 1780.689 \times NO_2^- + 4.885 \times NO_3^- + 3.774 \times NH_4^+ - 1.195 \times SUVA + 0.556 \times Dose - 0.021 \times CT$$

$$\mathbf{BAcAm} = 154.599 - 2775.487 \times UV_{254} - 31.619 \times DON + 8294.768 \times NO_2^- + 70.368 \times NO_3^- + 52.201 \times NH_4^+ + 3.821 \times SUVA + 1.157 \times Dose - 0.005 \times CT$$

$$\mathbf{HAcAms} = 130.369 - 2917.310 \times UV_{254} - 23.332 \times DON + 10\,725.382 \times NO_2^- + 77.140 \times NO_3^- + 67.713 \times NH_4^+ + 0.693 \times SUVA + 2.177 \times Dose + 0.035 \times CT$$

$$\mathbf{TCNM} = 98.194 - 1231.594 \times UV_{254} - 34.816 \times DOC + 4588.153 \times NO_2^- + 59.269 \times NO_3^- + 18.053 \times NH_4^+ + 1.028 \times SUVA + 0.524 \times Dose + 0.009 \times CT$$

$$\mathbf{DCP} = 123.067 - 3014.621 \times UV_{254} + 5.082 \times DOC + 18.328 \times NO_3^- - 29.578 \times NH_4^+ + 27.493 \times SUVA + 0.229 \times Dose + 0.012 \times CT$$

$$\mathbf{TCP} = 172.829 - 1766.448 \times UV_{254} - 36.097 \times DOC + 1541.053 \times NO_2^- + 43.800 \times NO_3^- - 23.634 \times NH_4^+ + 0.636 \times SUVA + 0.411 \times Dose + 0.022 \times CT$$

$$\mathbf{HKs} = 333.685 - 3513.762 \times UV_{254} - 50.565 \times DOC + 2023.126 \times NO_2^- + 44.713 \times NO_3^- - 30.395 \times NH_4^+ - 1.002 \times SUVA + 0.725 \times Dose + 0.035 \times CT$$

Table 3 summarizes the performance characteristics of the linear models for eDBPs prediction. The linear regression coefficients R^2 of the DBPs in water were all greater than 0.5. The F coefficient values were high, and this is important as a higher F value reflects a more significant impact of the predictor.[15] At the same time, the probability values p were <0.001, indicating that these models can be used to predict the formation of target DBPs.

Fig. 3 shows the correlation between experimentally determined and predicted concentrations of eDBPs. A straight line represents a 1 : 1 correlation. The Bland–Altman plots were used to graphically illustrate the agreement between experimentally determined concentrations and those predicted by the models (Fig. 4). A good agreement is shown in every case, as 95% of the points fall between the upper and lower limits. In addition, Fig. S1 shows the range of relative errors obtained for each compound. For all compounds, the average relative error was ≤30%, in an acceptable range, considering that these data represent different climatic impacts on source water. According to the non-parametric Wilcoxon test, there was no significant difference ($p > 0.05$) between the experimental concentrations and the calculated concentrations using the models.

Table 3 Performance evaluation of the linear models for the prediction of eDBPs

DBPs	R^2	F statistic	Probability p	Standard error
TCAN	0.782	5.570	0.000	0.319
DCAN	0.706	4.709	0.000	4.370
BCAN	0.779	18.767	0.000	5.400
DBAN	0.940	87.131	0.000	2.201
HANs	0.671	9.934	0.000	5.921
CAcAm	0.872	10.707	0.000	2.785
DCAcAm	0.854	7.277	0.000	3.494
BAcAm	0.797	21.071	0.000	7.611
HAcAms	0.824	25.585	0.000	9.601
TCNM	0.843	29.878	0.000	3.414
DCP	0.735	13.672	0.000	1.659
TCP	0.760	16.630	0.000	1.975
HKs	0.841	29.376	0.000	2.510

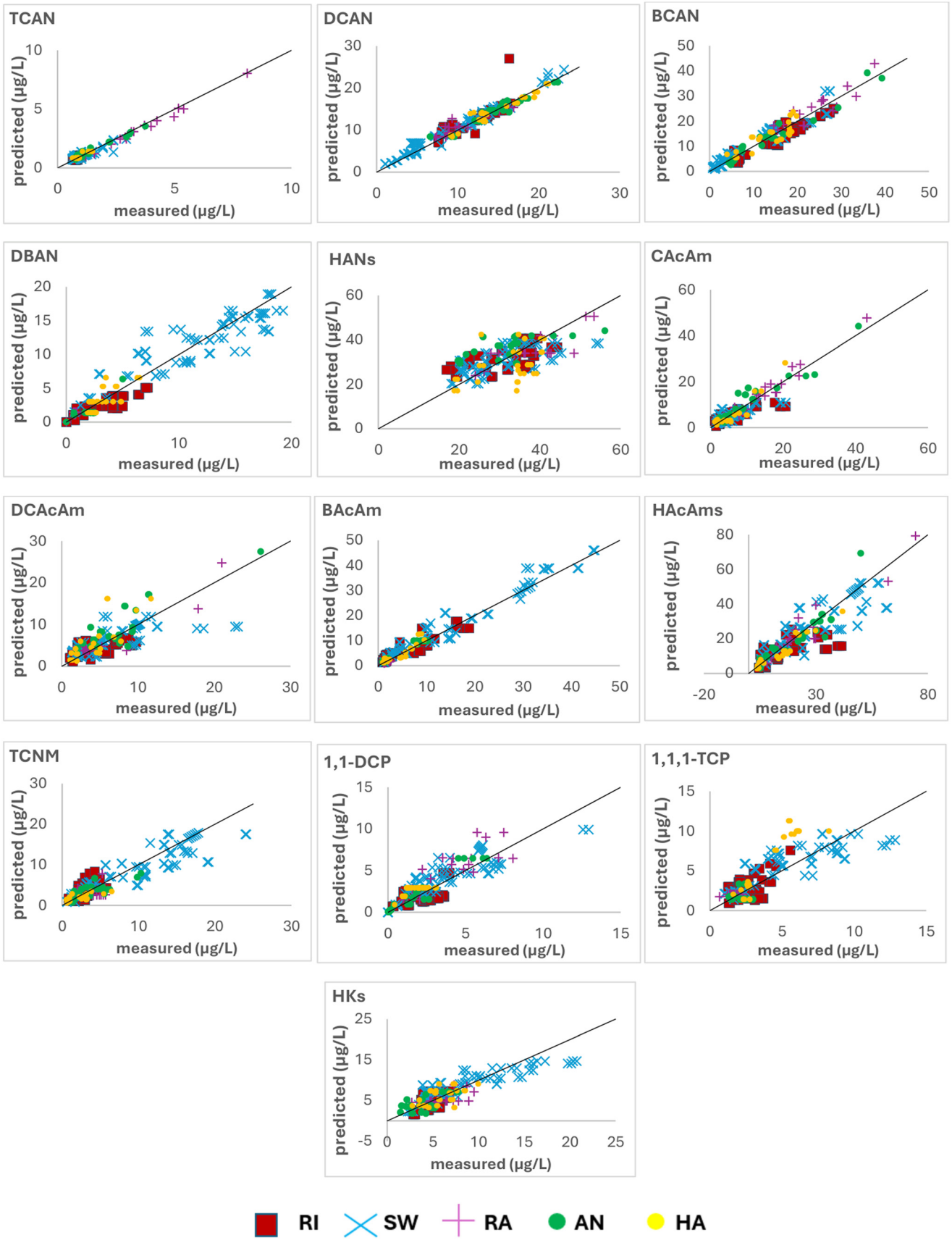

Fig. 3 Correlation between experimentally determined concentrations of eDBPs and values calculated based on linear models (RI: river sample, SW, RA, AN, HA represent river samples impacted by seawater, rainwater, algal organic matter and humic acids, respectively. The line represents a 1 : 1 relationship).

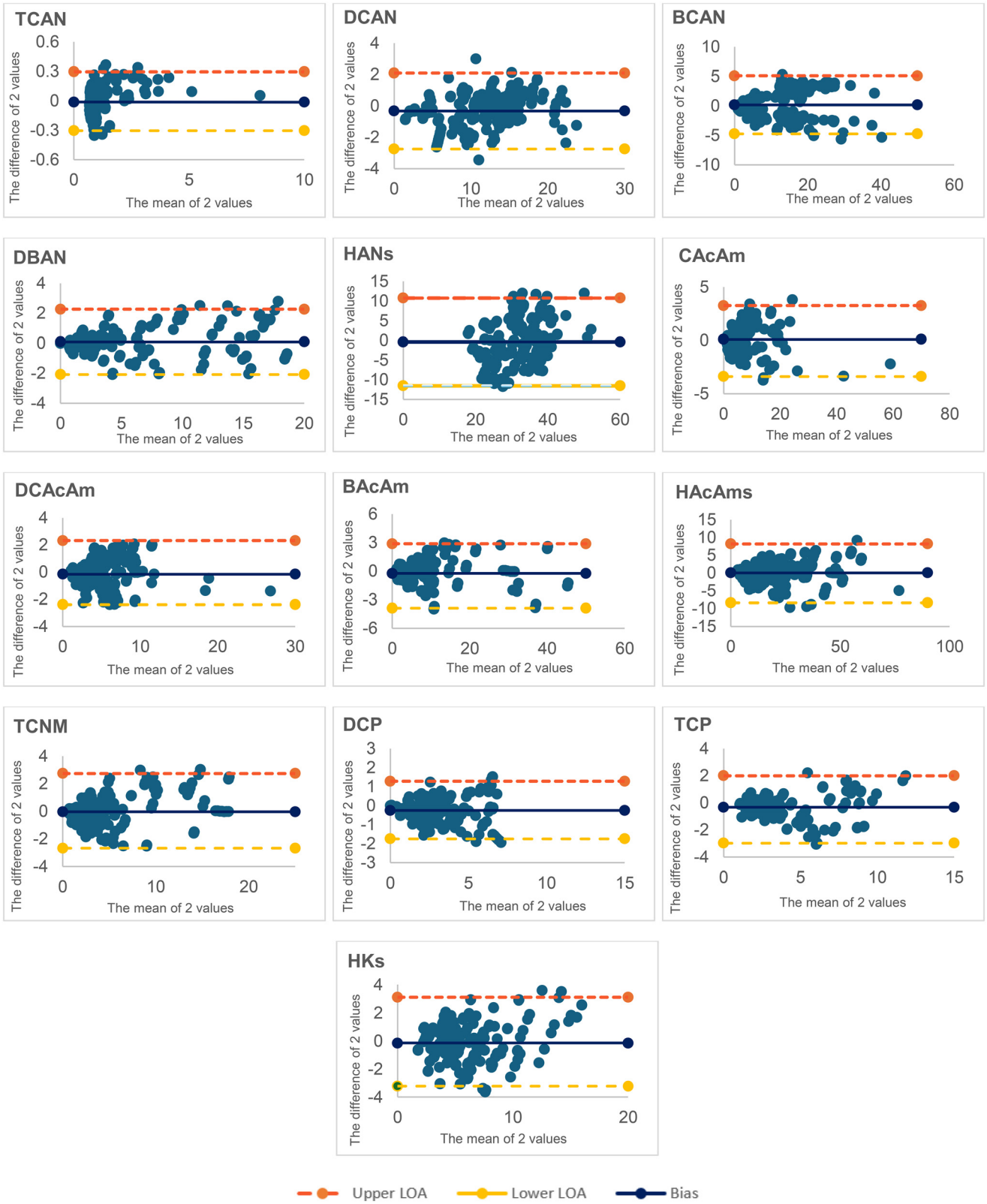

Fig. 4 Bland–Altman plots for the evaluation of the agreement between experimentally determined concentrations and values calculated based on linear models. The mean difference line represents the average bias between the two methods, while the upper and lower LOA are the limits of agreement (LOA) (mean difference ± 1.96 × standard deviation). Approximately 95% of data points should fall between these upper and lower limits to have a good agreement.

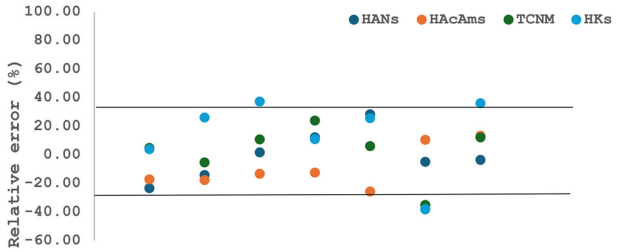

Fig. 5 Relative error between eDBP concentrations and values calculated based on linear models (external validation). Solid lines refer to 30% accepted error.

Finally, external validation was attempted in order to assess the reliability of the proposed models using experimental data that had not been included in the initial dataset for regression analysis. Fig. 5 shows the relative error between the concentrations of eDBPs and the predicted values calculated based on the linear models. Most samples were in an acceptable range, exhibiting a relative error of ≤30% for all groups of eDBPs.

In all the above proposed models, forms of nitrogen appear to play a major role in the formation of target eDBPs. While UV_{254}, SUVA and DOC are known as significant parameters, the inorganic forms of nitrogen were also major independent parameters in the formation of all the different groups, including HANs, HAcAms, TCNM and HKs. The significance of nitrogen forms is also referred to in the proposed models of Pang *et al.*, 2022,[13] in which NH_3 plays an important role in the formation of HANs, while Deng *et al.*, 2021,[41] indicated that NH_4^+ and NO_2^- are significant parameters in the formation of HKs. At the same time, Ersan *et al.* (2025)[42] included DON/DOC as a possible parameter for the prediction of HANs, whereas in our study, DON appears to contribute to the formation of HAcAms.

3.2. Logarithmic linear models

The efficiency of logarithmic models to predict eDBPs was also examined. The simplest way to model a nonlinear relationship is to transform the dependent variable y and/or the predictor variable x before estimating a regression model. The most commonly used transformation is the natural logarithm (ln). In this case, the values of dependent and independent variables were logarithmized before estimating the regression models, so the resulting prediction models have a logarithmic–logarithmic functional form. The resulting relationships are of the form:

$$\ln(y) = \text{constant} + a\ln(d) + b \times \ln(e) + c \times \ln(f) + \ldots,$$

where d, e, and f are the independent variables and a, b, and c are the regression coefficients.

The resulting logarithmic linear predictive models for each compound and class of DBPs are

$$\ln(\text{TCAN}) = -10.613 - 3.678 \times \ln(UV_{254}) - 0.263 \times \ln(NO_3^-) + 0.001 \times \ln(NH_4^+) - 0.262 \times \ln(\text{TN}) + 0.147 \times \ln(\text{SUVA}) + 0.409 \times \ln(\text{Dose}) - 0.047 \times \ln(\text{CT})$$

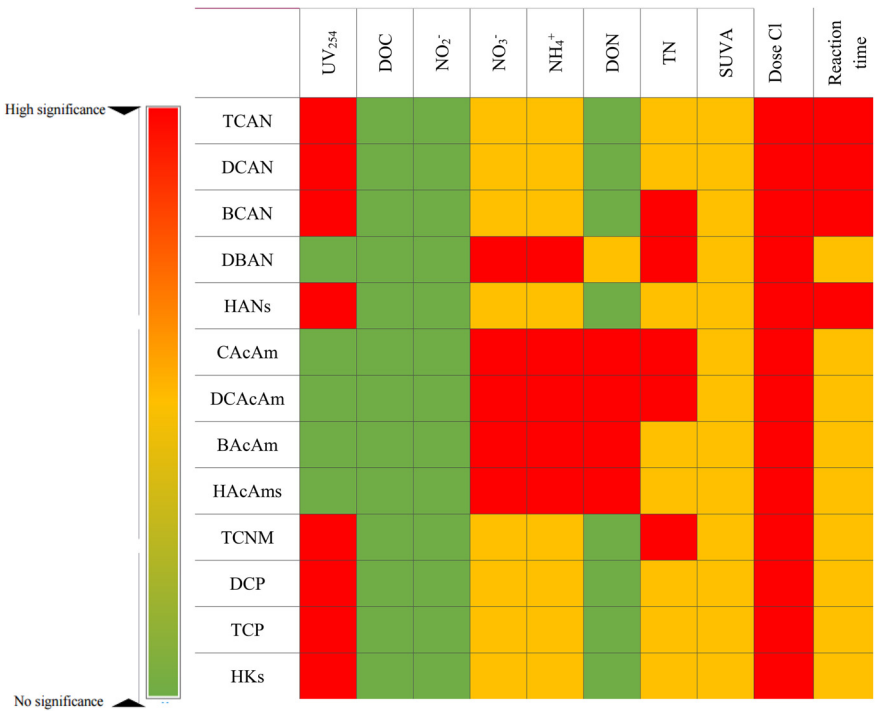

Fig. 6 Significance of the correlation between independent and dependent variables in logarithmic linear models. The parameters that presented no statistically significant correlations were excluded during the development of models.

$$\ln(\mathrm{DCAN}) = -8.627 - 3.982 \times \ln(\mathrm{UV_{254}}) + 0.225 \times \ln(\mathrm{NO_3^-}) - 0.139 \times \ln(\mathrm{NH_4^+}) - 0.436 \times \ln(\mathrm{TN}) - 0.785 \times \ln(\mathrm{SUVA}) + 0.272 \times \ln(\mathrm{Dose}) + 0.004 \times \ln(\mathrm{CT})$$

$$\ln(\mathrm{BCAN}) = -40.013 - 13.550 \times \ln(\mathrm{UV_{254}}) - 0.220 \times \ln(\mathrm{NO_3^-}) + 0.049 \times \ln(\mathrm{NH_4^+}) - 0.386 \times \ln(\mathrm{TN}) + 3.025 \times \ln(\mathrm{SUVA}) + 0.820 \times \ln(\mathrm{Dose}) + 0.260 \times \ln(\mathrm{CT})$$

$$\ln(\mathrm{DBAN}) = 2.092 + 2.979 \times \ln(\mathrm{NO_3^-}) - 0.100 \times \ln(\mathrm{NH_4^+}) - 0.098 \times \ln(\mathrm{DON}) - 0.578 \times \ln(\mathrm{TN}) + 0.386 \times \ln(\mathrm{SUVA}) + 0.400 \times \ln(\mathrm{Dose}) - 0.031 \times \ln(\mathrm{CT})$$

$$\ln(\mathrm{HANs}) = -2.469 - 2.002 \times \ln(\mathrm{UV_{254}}) - 0.231 \times \ln(\mathrm{NO_3^-}) + 0.017 \times \ln(\mathrm{NH_4^+}) - 0.252 \times \ln(\mathrm{TN}) - 0.587 \times \ln(\mathrm{SUVA}) + 0.397 \times \ln(\mathrm{Dose}) + 0.105 \times \ln(\mathrm{CT})$$

$$\ln(\mathrm{CAcAm}) = -34.577 - 0.981 \times \ln(\mathrm{NO_3^-}) + 0.039 \times \ln(\mathrm{NH_4^+}) - 12.287 \times \ln(\mathrm{DON}) + 0.688 \times \ln(\mathrm{TN}) + 0.424 \times \ln(\mathrm{SUVA}) + 0.679 \times \ln(\mathrm{Dose}) + 0.404 \times \ln(\mathrm{CT})$$

$$\ln(\mathrm{DCAcAm}) = -25.845 - 0.572 \times \ln(\mathrm{NO_3^-}) - 0.202 \times \ln(\mathrm{NH_4^+}) - 8.303 \times \ln(\mathrm{DON}) + 1.620 \times \ln(\mathrm{TN}) + 2.746 \times \ln(\mathrm{SUVA}) + 0.850 \times \ln(\mathrm{Dose}) - 0.061 \times \ln(\mathrm{CT})$$

$$\ln(\mathrm{BAcAm}) = -35.168 + 1.108 \times \ln(\mathrm{NO_3^-}) - 0.191 \times \ln(\mathrm{NH_4^+}) - 12.753 \times \ln(\mathrm{DON}) + 1.259 \times \ln(\mathrm{TN}) + 1.752 \times \ln(\mathrm{SUVA}) + 0.777 \times \ln(\mathrm{Dose}) - 0.010 \times \ln(\mathrm{CT})$$

$$\ln(\mathrm{HAcAms}) = -21.730 - 1.118 \times \ln(\mathrm{NO_3^-}) - 0.053 \times \ln(\mathrm{NH_4^+}) + 8.430 \times \ln(\mathrm{DON}) - 1.224 \times \ln(\mathrm{TN}) - 0.307 \times \ln(\mathrm{SUVA}) + 0.792 \times \ln(\mathrm{Dose}) + 0.147 \times \ln(\mathrm{CT})$$

$$\ln(\mathrm{TCNM}) = 3.829 + 1.323 \times \ln(\mathrm{UV_{254}}) - 0.341 \times \ln(\mathrm{NO_3^-}) + 0.108 \times \ln(\mathrm{NH_4^+}) + 0.490 \times \ln(\mathrm{TN}) + 0.034 \times \ln(\mathrm{SUVA}) + 0.745 \times \ln(\mathrm{Dose}) + 0.131 \times \ln(\mathrm{CT})$$

$$\ln(\mathrm{DCP}) = -22.799 - 1.151 \times \ln(\mathrm{UV_{254}}) + 9.367 \times \ln(\mathrm{NO_3^-}) - 0.280 \times \ln(\mathrm{NH_4^+}) + 1.466 \times \ln(\mathrm{TN}) + 22.106 \times \ln(\mathrm{SUVA}) + 0.716 \times \ln(\mathrm{Dose}) + 0.046 \times \ln(\mathrm{CT})$$

$$\ln(\mathrm{TCP}) = -12.683 - 3.236 \times \ln(\mathrm{UV_{254}}) + 4.288 \times \ln(\mathrm{NO_3^-}) + 0.043 \times \ln(\mathrm{NH_4^+}) - 1.513 \times \ln(\mathrm{TN}) + 5.416 \times \ln(\mathrm{SUVA}) + 0.520 \times \ln(\mathrm{Dose}) + 0.226 \times \ln(\mathrm{CT})$$

$$\ln(\mathrm{HKs}) = -14.260 - 4.567 \times \ln(\mathrm{UV_{254}}) + 2.987 \times \ln(\mathrm{NO_3^-}) + 0.003 \times \ln(\mathrm{NH_4^+}) - 1.529 \times \ln(\mathrm{TN}) + 3.682 \times \ln(\mathrm{SUVA}) + 0.605 \times \ln(\mathrm{Dose}) + 0.146 \times \ln(\mathrm{CT})$$

Table 4 Performance evaluation of the logarithmic linear models for the prediction of eDBPs

DBPs	R^2	F statistic	Probability p	Standard error
TCAN	0.863	5.387	<0.001	0.262
DCAN	0.841	5.354	<0.001	0.544
BCAN	0.791	20.206	<0.001	0.657
DBAN	0.964	149.382	<0.001	0.307
HANs	0.680	10.429	<0.001	0.192
CAcAm	0.665	9.634	<0.001	0.481
DCAcAm	0.861	7.835	<0.001	0.568
BAcAm	0.856	33.278	<0.001	0.640
HAcAms	0.803	21.954	<0.001	0.475
TCNM	0.826	25.991	<0.001	0.522
DCP	0.704	11.415	<0.001	0.564
TCP	0.747	15.340	<0.001	0.450
HKs	0.864	35.779	<0.001	0.276

Table 4 summarizes the performance characteristics of the ln-linear models for eDBPs prediction. The ln-linear regression coefficients R^2 of the DBPs in water were all greater than 0.5. The F-statistic values were also high, and the p-values were all <0.001, indicating that these models can also be used to predict the formation of these specific DBPs.

Fig. 7 shows the correlation between the logarithms of the actual concentrations and the calculated logarithmic values. The straight line represents a 1 : 1 correlation. In addition, Fig. S2 shows the range of relative errors obtained for each compound. For all compounds, the mean relative error was ≤30%, in an acceptable range. Fig. 8 presents the Bland–Altman plots. They also show good agreement in every case, as 95% of the points fall between the upper and lower limits of agreement. According to the non-parametric Wilcoxon test, there was no significant difference ($p > 0.05$) between the experimental concentrations and the calculated concentrations from the models.

Fig. 9 shows the relative error between the concentrations of eDBPs and the predicted values calculated using the logarithmic linear models. The reliability of the above models was checked using experimental data that had not been used in the regression analysis set (external validation). Almost all samples showed a relative error of ≤30% for all compounds.

4. Conclusions

This study aimed to develop and evaluate predictive relationships between emerging DBPs and routinely monitored water quality parameters and chlor(am)ination conditions. 221 samples were used as a dataset for the development of two kinds of multivariate predictive models: one linear and one logarithmic (ln-linear). The proposed models for each compound and for each class of compounds were derived from multiple regression analysis. In both cases, the proposed relationships incorporate forms of nitrogen as

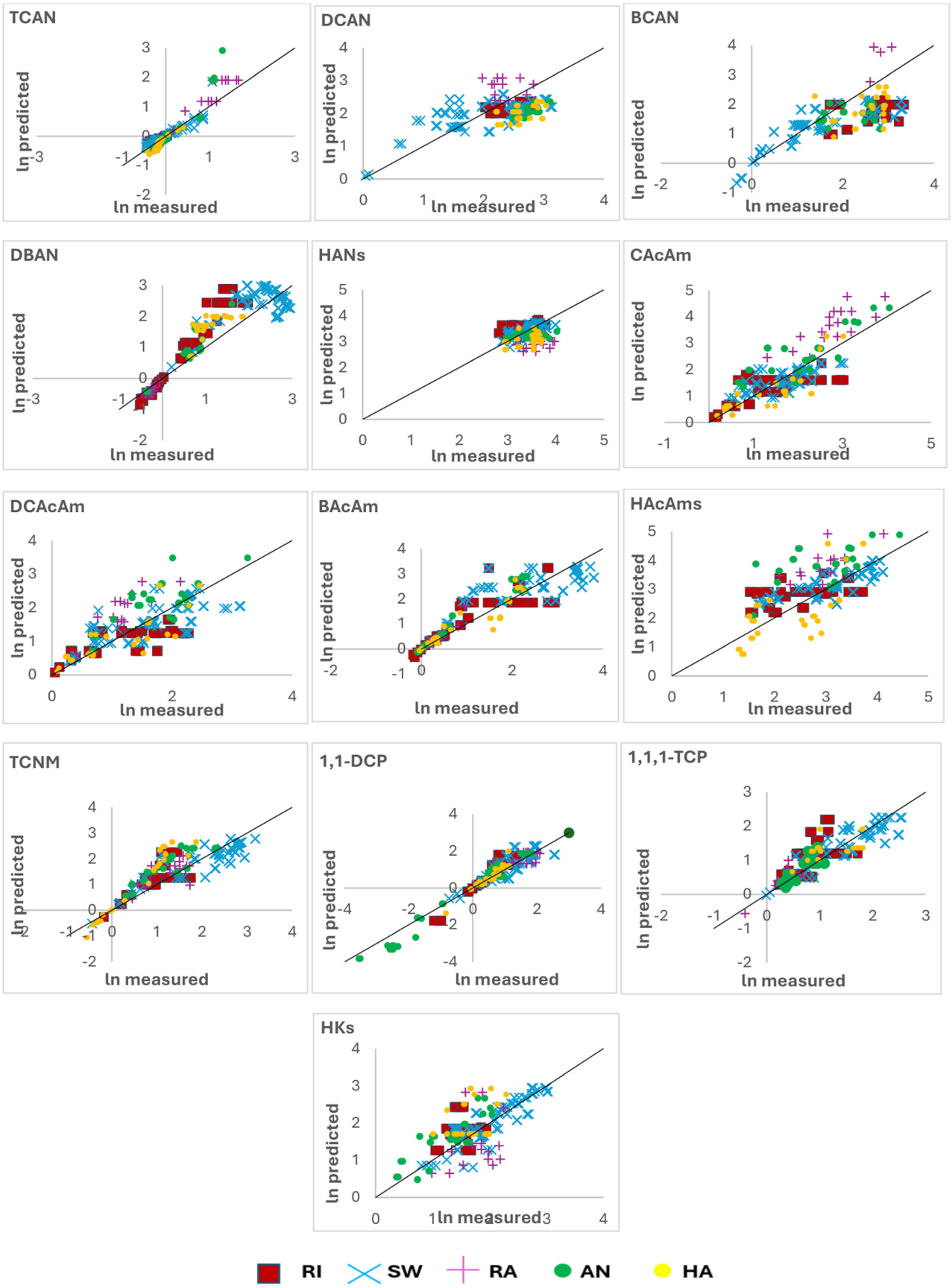

Fig. 7 Correlation between experimentally determined eDBP concentrations and values calculated based on logarithmic linear models (RI: river sample; SW, RA, AN, and HA represent river samples impacted by seawater, rainwater, algal organic matter and humic acids, respectively. The line represents a 1 : 1 relationship).

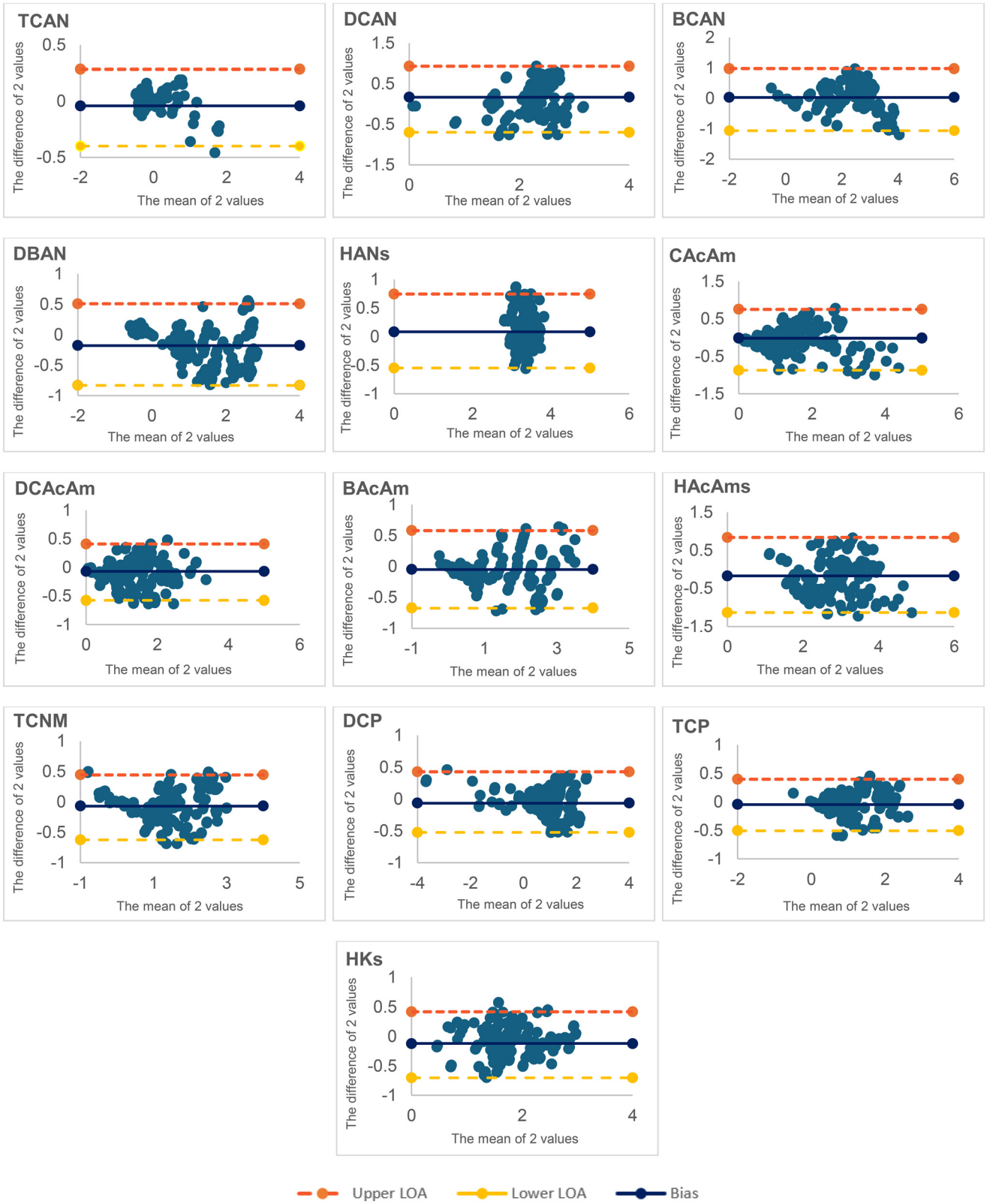

Fig. 8 Bland–Altman plots for the evaluation of agreement between experimentally determined concentrations and values calculated based on logarithmic linear models. The mean difference line represents the average bias between the two methods, while the upper and lower LOAs are the limits of agreement (LOA) (mean difference ± 1.96 × standard deviation). Approximately 95% of data points should fall between these upper and lower limits to have a good agreement.

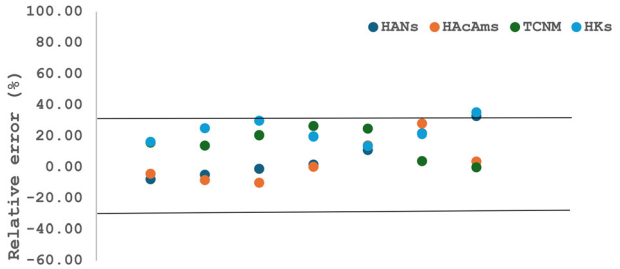

Fig. 9 Relative error between eDBP concentrations and values calculated based on linear logarithmic models (external validation). Solid lines refer to 30% accepted error.

significant predictors for the formation of eDBPs. At the same time, disinfection contact time and disinfectant dose were also significant in the formation of eDBPs. The proposed models presented an average range of relative error ≤30% for each compound. Also, using the Bland–Altman method for the evaluation of agreement between experimentally determined concentrations and values calculated based on the models, good agreement was observed for both linear and logarithmic models, indicating that they can be used to predict the formation of target DBPs. There was no statistically significant difference between experimental and calculated concentrations of DBPs. The compounds TCAN, DCAN, BCAN, DBAN, CAcAm, DCAcAm, and BAcAm presented slightly better correspondence to the linear models, while the compounds TCNM, DCP, TCP, and the set of HKs showed better correspondence to the logarithmic linear models. The group of HANs and HAcAms presented the same performance in both models.

The approach and findings of this study are useful to evaluate exposure to non-monitored DBPs, although they appear to be more toxic than the regulated compounds. These predictive relationships can be a useful tool for water managers and regulators of drinking water for the identification and control of the occurrence of eDBPs and evaluation of exposure risks across a range of water variations due to climate change.

Author contributions

Argyri Kozari: investigation, methodology, validation, visualization, data curation, writing – original draft. Dimitra Voutsa: conceptualization, methodology, resources, investigation, writing – review & editing, supervision.

Conflicts of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this study.

Data availability

The data used in this article were obtained experimentally and have been published at Kozari *et al.*, 2024 (DOI: https://doi.org/10.1007/s11356-024-32960-4) and Kozari and Voutsa, 2023 (DOI: https://doi.org/10.1016/j.scitotenv.2023.166041).

Supplementary information (SI): the supplementary file contains figures that present the relative errors between measured and predicted concentrations of eDBPs from the linear and logarithmic models. See DOI: https://doi.org/10.1039/d5ew01176k.

Acknowledgements

This research was co-financed by Greece and the European Union (European Social Fund-ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning" in the context of the Act "Enhancing Human Resources Research Potential by undertaking a Doctoral Research" Sub-action 2: IKY Scholarship Programme for PhD candidates in the Greek Universities".

References

1 Directive (EU) 2020/2184 of the Euroean Parliament and of the Council of 16 December 2020 on the quality of water intended for human consumption (recast). Official Journal L 435, 23.12.2020, pp. 1–62.
2 S. D. Richardson, M. J. Plewa, E. D. Wagner, R. Schoeny and D. M. DeMarini, Occurrence, genotoxicity and carcinogenicity of regulated and emerging disinfection by-products in drinking water: A review and roadmap for research, *Mutat. Res.*, 2007, **363**(1–3), 178–242.
3 P. M. Forster, C. Smith, T. Walsh, W. F. Lamb, R. Lamboll, C. Cassou, M. Hauser, Z. Hausfather, J. Y. Lee, M. D. Palmen, K. von Schuckmann, A. B. A. Slangen, S. Szopa, B. Trewin, J. Yun, N. D. Gillett, S. Jenkins, H. D. Matthews, K. Raghavan, A. Ribes, J. Rogelj, D. Rosen, X. Zhang, M. Allen, L. A. Reis, R. M. Andrew, R. A. Betts, A. Borger, J. A. Broersma, S. N. Burgess, L. Cheng, P. Freidlingstein, C. M. Domingues, M. Gambarini, T. Gasser, J. Gütschow, M. Ishii, C. Kadow, J. Kennedy, R. E. Killick, P. B. Krummel, A. Liné, D. P. Monselesan, C. Morice, J. Mühle, V. Naik, G. P. Peters, A. Pirani, J. Pongratz, J. C. Minx, M. Rigby, R. Rohde, A. Savita, S. I. Seneviratne, P. Thorne, C. Wells, L. M. Western, G. R. van der Werf, S. E. Wijffels, V. Masson-Delmotte and P. Zhai, Indicators of Global Change 2024: annual update of key indicators of the state of the climate system and human influence, *Earth Syst. Sci. Data*, 2025, **17**(6), 2641–2680.
4 R. Sadiq, M. J. Rodriguez and H. R. Mian, Empirical Models to predict disinfection by-products (DBPs) in drinking water: An updated review, *Encyclop. of Environ. Health*, Elsevier, 2nd edn, 2019, pp. 324–338.
5 T. Bond, M. R. Templeton, N. H. M. Kamal, N. Graham and R. Kanda, Nitrogenous disinfection byproducts in English drinking water supply systems: Occurrence, bromine substitution and correlation analysis, *Water Res.*, 2015, **85**, 85–94.
6 C. Postigo, P. Emiliano, D. Barceló and F. Valero, Chemical characterization and relative toxicity assessment of

disinfection byproduct mixtures in a large drinking water supply network, *J. Hazard. Mater.*, 2018, **359**, 166–173.

7 M. J. Plewa, E. D. Wagner and S. D. Richardson, TIC-TOX: A preliminary discussion on identifying the forcing agents of DBP-mediated toxicity of disinfected water, *J. Environ. Sci.*, 2017, **58**, 208–216.

8 E. D. Wagner and M. J. Plewa, CHO cell cytotoxicity and genotoxicity analyses of -disinfection by-products: An updated review, *J. Environ. Sci.*, 2017, **58**, 64–76.

9 S. W. Krasner, A. Jia, C. F. T. Lee, R. Shirkhani, J. M. Allen, S. D. Richardson and M. J. Plewa, Relationships between regulated DBPs and emerging DBPs of health concern in U. S. drinking water, *J. Environ. Sci.*, 2022, **117**, 161–172.

10 I. Kalita, A. Kamilaris, P. Havinga and I. Reva, Assessing the Health Impact of Disinfection Byproducts in Drinking Water, *ACS ES&T Water*, 2024, **4**(4), 1564–1578.

11 A. Kozari, S. Gkelis and D. Voutsa, Impact of climate change on formation of nitrogenous disinfection by-products. Part II: water blooming and enrichment by humic substances, *Environ. Sci. Pollut. Res.*, 2024, DOI: 10.1007/s11356-024-32960-4.

12 A. Kozari and D. Voutsa, Impact of climate change on formation of nitrogenous disinfection by products. Part I: Sea level rise and flooding events, *Sci. Total Environ.*, 2023, **901**, 166041.

13 Z. Pang, P. Zhang, X. Chen, F. Dong, J. Deng, C. Li, J. Liu, X. Ma and A. M. Dietrich, Occurrence and modeling of disinfection byproducts in distributed water of a megacity in China: Implications for human health, *Sci. Total Environ.*, 2022, **848**, 157674.

14 H. Hong, Z. Zhang, A. Guo, L. Shen, H. Sun and Y. Liang, Radial basis function artificial neural network (RBF ANN) as well as the hybrid method of RBF ANN and grey relational analysis able to well predict trihalomethanes levels in tap water, *J. Hydrol.*, 2020, **591**, 125574.

15 H. R. Mian, G. Chhipi-Shrestha, K. Hewage, M. J. Rodriguez and R. Sadiq, Predicting unregulated disinfection by-products in small water distribution networks: an empirical modelling framework, *Environ. Monit. Assess.*, 2020, **192**, 497.

16 B. Ye, W. Wang, L. Yang and J. E. X. Wei, Formation and modeling of disinfection byproducts in drinking water of six cities in China, *J. Environ. Monit.*, 2011, **13**, 1271–1275.

17 S. Chowdhury, P. Champagne and P. J. McLellan, Models for predicting disinfection byproduct (DBP) formation in drinking waters: a chronological review, *Sci. Total Environ.*, 2009, **407**, 4189–4206.

18 L. Liang and P. C. Singer, Factors influencing the formation and relative distribution of haloacetic acids and trihalomethanes in drinking water, *Environ. Sci. Technol.*, 2003, **37**, 2920–2928.

19 P. C. Singer and K. Bilyk, Enhanced coagulation using a magnetic ion exchange re-sin, *Water Res.*, 2002, **36**, 4009–4022.

20 Y. Ke, W. Sun, Z. Chu, Y. Zhu, X. Chen, S. Yan, Y. Li and S. Xie, Effects of disinfectant type and dosage on biofilm's activity, viability, microbiome and antibiotic resistome in bench-scale drinking water distribution systems, *Water Res.*, 2024, **249**, 120958.

21 Y. Liang, R. Huang, J. Wang, Z. Han, S. Wu, Y. Tan, X. Huangfu and Q. He, Machine learning-guided prediction of chlorinated/chloraminated disinfection by-product formation in drinking water treatment, *Water Res.*, 2025, **283**, 123849.

22 L. Godo-Pla, J. Rodríguez, J. Suquet, P. Emiliano, F. Valero, M. Pocha and H. Monclús, Control of primary disinfection in a drinking water treatment plantbased on a fuzzy inference system, *Process Saf. Environ. Prot.*, 2021, **145**, 63–70.

23 G. Hu, H. R. Mian, S. Mohammadium, M. Rodrigues, K. Hewage and R. Sadiq, Appraisal of machine learning techniques for predicting emerging disinfection byproducts in small water distribution networks, *J. Hazard. Mater.*, 2023, **446**, 130633.

24 P. Kulkarni and S. Chellam, Disinfection by-product formation following chlorination of drinking water: artificial neural network models and changes in speciation with treatment, *Sci. Total Environ.*, 2010, **408**(19), 4202–4210.

25 R. A. Li, J. A. McDonald, A. Sathasivan and S. J. Khan, A multivariate Bayesian network analysis of water quality factors influencing trihalomethanes formation in drinking water distribution systems, *Water Res.*, 2021, **190**, 116712.

26 J. K. Mahato and S. K. Gupta, Advanced oxidation of Trihalomethane (THMs) precursors and season-wise multi-pathway human carcinogenic risk assessment in Indian drinking water supplies, *Process Saf. Environ. Prot.*, 2022, **159**, 996–1007.

27 J. Park, C. H. Lee, K. H. Cho, S. Hong, Y. M. Kim and Y. Park, Modeling trihalomethanes concentrations in water treatment plants using machine learning techniques, *Desalin. Water Treat.*, 2018, **111**, 125–133.

28 F. Peng, Y. Lu, Y. Wang, L. Yang, Z. Yang and H. Li, Predicting the formation of disinfection by-products using multiple linear and machine learning regression, *J. Environ. Chem. Eng.*, 2023, **11**, 110612.

29 R. Sikder, T. Zhang and T. Ye, Predicting THM Formation and revealing its contributors in drinking water treatment using machine learning, *ACS ES&T Water*, 2023, **4**(3), 899–912.

30 K. P. Singh and S. Gupta, Artificial intelligence based modeling for predicting the disinfection by-products in water, *Chemom. Intell. Lab. Syst.*, 2012, **114**, 122–131.

31 J. Liu, L. Ling, Q. Hu, C. Wang and C. Shang, Effects of operating conditions on disinfection by-product formation, calculated toxicity, and changes in organic matter structures during seawater chlorination, *Water Res.*, 2022, **220**, 118631.

32 C. J. Chang, C. P. Huang, C. Y. Chen and G. S. Wang, Assessing the potential effect of extreme weather on water quality and disinfection by-product formation using laboratory simulation, *Water Res.*, 2020, **170**, 115296.

33 L. Li, Y. Li, Z. Fang and C. He, Study on molecular structure characteristics of natural dissolved organic nitrogen by use of negative and positive ion mode electrospray ionization

Orbitrap mass spectrometry and collision-induced dissociation, *Sci. Total Environ.*, 2022, **810**, 152116.

34 W. Chu, X. Li, T. Bond, N. Gao and D. Yin, The formation of haloacetamides and other disinfection by-products from non-nitrogenous low-molecular weight organic acids during chloramination, *J. Chem. Eng.*, 2016, **285**, 164–171.

35 H. Hong, Y. Xiaoqing, X. Song, Y. Qin, H. Sun, H. Lin, J. Chen and Y. Liang, Bromine incorporation into five DBP classes upon chlorination of water with extremely low SUVA values, *Sci. Total Environ.*, 2017, **22**, 156–166.

36 A. Kanan and T. Karanfil, Estimation of haloacetonitriles formation in water: Uniform formation conditions versus formation potential tests, *Sci. Total Environ.*, 2020, **744**, 140987.

37 C. Sfynia, T. Bond, N. Ganidi, R. Kanda and M. R. Templeton, Predicting the formation of haloacetonitriles and haloacetamides by simulated distribution system tests, *Procedia Eng.*, 2017, **186**, 186–192.

38 A. Carratalà, V. Bachmann, T. R. Julian and T. Kohn, Adaptation of human Enterovirus to warm environments leads to resistance against chlorine disinfection, *Environ. Sci. Technol.*, 2020, **54**, 11292–11300.

39 J. Li, S. Ren, X. Qiu, S. Zhao, R. Wang and Y. Wang, Electroactive ultrafiltration membrane for simultaneous removal of antibiotic, antibiotic resistant bacteria, and antibiotic resistance genes from wastewater effluent, *Environ. Sci. Technol.*, 2022, **56**, 15120–15129.

40 A. Ruecker, H. Uzun, T. Karanfil, M. T. K. Tsui and A. T. Chow, Disinfection byproduct precursor dynamics and water treatability during an extreme flooding event in a coastal blackwater river in Southeastern United States, *Chemosphere*, 2017, **188**, 90–98.

41 Y. Deng, X. Zhoua, J. Shen, G. Xiao, Y. Hong, H. Lin, F. Wub and B. Q. Liaoc, New methods based on back propagation (BP) and radial basis function (RBF) artificial neural networks (ANNs) for predicting the occurrence of haloketones in tap water, *Sci. Total Environ.*, 2021, **772**, 145534.

42 G. Ersan, E. Goz and T. Karanfil, Performance analysis of machine learning algorithms for the prediction of disinfection byproducts formation during chlorination: Effect of background water characteristics, *J. Environ. Manag.*, 2025, **389**, 126144.

43 G. Chhipi-Shrestha, M. Rodriguez and R. Sadiq, Framework for cost-effective prediction of unregulated dis-infection by-products in drinking water distribution using differential free chlorine, *Environ. Sci.: Water Res. Technol.*, 2018, **4**(10), 1564–1576.

44 S. Chowdhury, K. A. Sattar and S. M. Rahman, Predicting few disinfection byproducts in the water distribution systems using machine learning models, *Environ. Sci. Pollut. Res.*, 2025, **32**(7), 3776–3794.

45 S. Moradi, C. W. K. Chow, D. Cook, G. Newcombe and R. Amal, Estimating NDMA formation in a distribution system using a hybrid genetic algorithm, *J. - Am. Water Works Assoc.*, 2017, **109**(6), E265–E272.

46 C. M. Botha, S. C. Liebenberg, F. H. Conradie and A. F. van der Merwe, The potential of the Bland-Altman method in chemical engineering, *S. Afr. J. Chem. Eng.*, 2025, **54**, 348–356.

47 S. L. Zubaidi, K. Hashim, S. Ethaib, N. S. S. Al-Bdairi, H. Al-Bugharbee and S. K. Gharghan, A novel methodology to predict monthly municipal water demand based on weather variables scenario, *J. King Saud Univ., Eng. Sci.*, 2022, **34**(3), 163–169.