







<https://doi.org/10.1590/2318-0331.282320220118>

## Data imputation of water quality parameters through feed-forward neural networks

### *Preenchimento de dados de parâmetros de qualidade da água por redes neurais artificiais*

Luis Otávio Miranda Peixoto<sup>1</sup> , Bárbara Alves de Lima<sup>1</sup> , Camila de Carvalho Almeida<sup>1</sup> ,  
Cristóvão Vicente Scapulatempo Fernandes<sup>1</sup> , Jorge Antonio Silva Centeno<sup>1</sup>  & Júlio César Rodrigues de Azevedo<sup>1</sup> 

<sup>1</sup>Universidade Federal do Paraná, Curitiba, PR, Brasil

E-mails: luisotaviopexoto@gmail.com (LOMP), babi.lima90@gmail.com (BAL), cami.almeidac@gmail.com (CCA), cris.dhs@ufpr.br (CVSF), centeno@ufpr.br (JASC), jcrazevedo@hotmail.com (JCRA)

Received: December 18, 2022 - Revised: March 29, 2023 - Accepted: April 11, 2023

## ABSTRACT

The constant monitoring of water quality is fundamental for the understanding of the aquatic environment, yet it demands great financial investments and is susceptible to inconsistencies and missing values. Using a database composed of 59 sampling campaigns, performed for 12 years, on 10 monitoring stations along the Iguassu River Basin (Southern Brazil), this study presents a model, based on feed-forward neural networks, which imputed 1,370 values for 11 traditional water quality parameters, as well as 3 contaminants of emerging concern (caffeine, estradiol and ethinylestradiol). The model validation errors varied from 0.978 mg L<sup>-1</sup> and 0.017 mg L<sup>-1</sup> for the traditional parameters, for caffeine the validation error was of 0.212 µg L<sup>-1</sup> and for the hormones, the errors were of 0.04 µg L<sup>-1</sup> (E1) and 0.044 µg L<sup>-1</sup> (EE1). The models underwent two techniques to understand the operations performed within the model (isolation and nullification), which were consistent to those explained by natural processes. The results point to the validity of modeling water quality parameters (especially the concentrations of caffeine) through neural networks, which could lead to better resource allocation in environmental monitoring, as well as improving available datasets and valuing previous monitoring efforts.

**Keywords:** Contaminants of emerging concern; Artificial Intelligence; Environmental monitoring.

## RESUMO

O monitoramento constante da qualidade da água é fundamental para o entendimento dos ambientes aquáticos, mas este esforço demanda grandes investimentos financeiros, além de estar suscetível a inconsistências e falhas na obtenção dos valores. Usando uma base de dados composta por 59 campanhas amostrais, realizadas durante um período de 12 anos, em 10 estações de monitoramento ao longo da bacia do rio Iguaçu, este trabalho apresentou um modelo baseado em redes neurais artificiais que preencheu 1.370 valores para onze parâmetros tradicionais de qualidade da água. Os erros de validação do modelo variaram de 0,978 mg L<sup>-1</sup> a 0,017 mg L<sup>-1</sup> para os parâmetros tradicionais, enquanto para a cafeína, este erro foi de 0,212 µg L<sup>-1</sup> e para os hormônios de 0,04 µg L<sup>-1</sup> (E1) e 0,044 µg L<sup>-1</sup> (EE1). Duas técnicas (isolamento e anulação) foram aplicadas nos modelos para se entender o relacionamento treinado pelo modelo entre as variáveis de entrada e saída. Os resultados apontam para a viabilidade da aplicação de redes neurais para a modelagem de parâmetros de qualidade da água (em especial, a cafeína), o que poderia levar a uma melhor alocação de recursos no monitoramento ambiental, além de expandir os bancos de dados disponíveis e valorizar os esforços despendidos para este monitoramento.

**Palavras-chave:** Contaminantes emergentes; Inteligência artificial; Monitoramento ambiental.

## INTRODUCTION

Water is essential for human life but increasing urbanization and industrialization of modern society, as well as the ever-increasing agricultural activities demanded to support the populational rise, has put pressures onto water resources systems during the past century. Therefore, the continuous and systematic monitoring of water quality is a fundamental part of the understanding and preservation of economic, environmental, and human health.

Monitoring water quality of a large water body during extended periods of time demands large financial investments in laboratorial infrastructure, qualified personnel and sampling logistics (Woodhouse & Muller, 2017). Yet, these investments do not guarantee a flawless and complete dataset, as human error, instrumental malfunction, differences in methodology, budgetary constraints, and/or implementation of different monitoring approaches allow for missing or incongruent data collection within a studied period. These problems are more evident when dealing with concentration measurements for contaminants of emerging concern (Muthukrishnan et al., 2020).

Contaminants of emerging concern are natural or artificial compounds which are originated from human activity and are present in water resources in extremely low concentrations (in the order of ng to ug L<sup>-1</sup>), such as pharmaceuticals, female sex hormones, industrial preservatives, and additives, as well as agricultural pesticides, and personal care products (Berrou et al., 2021). The chronic exposure to such compounds, even at low concentrations, though not completely understood, have been linked to mutagenicity, carcinogenicity, endocrine disruption, deleterious alterations in reproductive patterns and environmental imbalance (Giri, 1993; Kidd et al., 2007; Oaks et al., 2004; Isidori et al., 2005; Katipoglu-Yazan et al., 2013; Galus et al., 2013).

The greatest hurdle that must be dealt with when dealing with water resources variables is the high variability and intertwined variance between the variables present. The relationships which govern the behavior of traditional parameters and contaminants of emerging concern within an aquatic environment (as well as interposed to the socioeconomical picture in which this system is enclosed) are extremely complex. One of the modern alternatives to understanding such complex databases and interrelationships is artificial intelligence, as computational modeling has experienced great advances on the past few decades. Machine learning methods, for example, are computational algorithms developed aiming at interpreting and reproducing intrinsic patterns in highly-dimensional datasets.

Several studies have been performed to analyze the viability of using artificial intelligence methods for predicting the concentrations of different chemical parameters within aquatic ecosystems, such as the biological oxygen demand (Banejad & Olyaie, 2011; Ahmed & Shah, 2017; Ahmadi et al., 2018; Ooi et al., 2022), particulate organic carbon (Buchard-Levine et al., 2014; Zhang et al., 2021), dissolved organic carbon (Buchard-Levine et al., 2014; Zhou, 2020), ammonia nitrogen (Suen & Eheart, 2003; Wang et al., 2013; Ahmed et al., 2019; Hayder et al., 2021), nitrate and nitrite (Kamyab-Talesh et al., 2019; Ha et al., 2020; Li et al., 2020a), total nitrogen and total phosphorus (Liu & Lu, 2014; Shen et al., 2019; Wang et al., 2021; Lu et al., 2022), orthophosphate (Ha et al.,

2020), and dissolved oxygen (Banejad & Olyaie, 2011; Abba et al., 2017; Csábrági et al., 2019; Li et al., 2020b; Ahmed et al., 2021).

There are very few studies that have analyzed the application of machine learning to model the concentration of contaminants of emerging concern (Kiesling et al., 2019; Krishnaraj & Deka, 2020), such as caffeine and female sex hormones. Also, very few studies have been performed on the practicality of data imputation for highly irregular, not evenly spaced temporally, water quality databases (Park et al., 2021).

Thus, this study aims at applying artificial intelligence models, based on feed-forward neural networks, to predict missing data within a water quality monitoring dataset produced for the Iguassu River from 2005 and 2017. The parameters that will be modeled are the concentrations of 11 traditional water quality parameters: biological oxygen, particulate organic carbon, dissolved organic carbon, ammonia nitrogen, nitrate, nitrite, total nitrogen, total phosphorus, and orthophosphate, as well as four contaminants of emerging concern: caffeine, estradiol, ethinylestradiol, and estrone. This study also aims at understanding the accuracy that may be expected in such an endeavor, as well as the intrinsic functions and relationships that each model was able to extract from the dataset for each pair of input-output variables. Thus, the models created to impute the missing data of the set might be used to predict values for faulty future sampling campaigns, as well as be able to better understand the natural system it is applied, which could lead to more efficient resource allocation for future environmental monitoring.

## MATERIALS AND METHODS

### Study area

The database of water quality parameters used for this study was collected from the Iguassu River, in the State of Paraná in Southern Brazil. It runs for 1.311 km, and its watershed spreads over 70,800 km<sup>2</sup>, 97% of it is located within Brazilian borders, while the remaining 3% is in Argentina. The study area does not cover the entirety of the Iguassu basin, being restricted to 39% of the basin area and 36% of the river's course.

This study analyzed data collected from 10 different sampling sites across the first half of the Iguassu River. Table 1 shows the coordinates and distance from the spring for each of the sampling

**Table 1.** Coordinates of the ten sampling sites on the Iguassu River.

Site	Coordinates		Distance from spring
	Latitude	Longitude	
IG1	49° 10' 12" W	25° 27' 1" S	22.5
IG2D/IG2E	49° 11' 24" W	25° 28' 48" S	30.4
IG3	49° 15' 36" W	25° 36' 0" S	47.5
IG4	49° 22' 12" W	25° 37' 48" S	67.5
IG5	49° 30' 36" W	25° 36' 2" S	84.9
IG6	49° 37' 48" W	25° 35' 24" S	107.2
IG7	49° 53' 24" W	25° 33' 0" S	154.5
IG8	50° 23' 24" W	25° 52' 48" S	293.2
IG9	51° 4' 48" W	26° 13' 48" S	473.1

sites. Figure 1 illustrates the location of the sampling sites on the Iguassu River Basin.

Though the Iguassu River is formed by the confluence of the Irai and Atuba River (around IG2D and IG2E), the Irai River's spring, located 22 kilometers upstream from IG1 (located in a mountainous Atlantic rainforest biome), is commonly addressed as the Iguassu River's spring. For this study, this will also be assumed.

The IG1 sampling site is in the municipality of Pinhais, within the Curitiba Metropolitan Area (CMA), which is the largest and most populous urban agglomeration of the State of Paraná and the 9<sup>th</sup> most populous of Brazil. This confers urban characteristics to the majority of the first 100 kilometers of the Iguassu River's course.

The sites IG2D and IG2E are located within the same transversal section of the river but represents different aspects since are located 50 meters downstream from the confluence of the Atuba and Irai rivers, while their waters have yet to completely mix. The characteristics of these water bodies are contrastingly different, as the Irai river is part of the water supply system of the region, it tends to present less degraded water quality (IG2E) than that present on the Atuba River (IG2D), which receives the discharge of one of the largest wastewater treatment plants of the CMA, usually degrading the water quality of the river.

The subbasins for IG3, IG4 and IG5 are encompassed entirely within an urban setting. The IG6 subbasin is situated in a semi-urban scenario. The first sampling site located in a mostly

rural environment is IG7, situated in the municipality of Porto Amazonas. The sub-basins of IG8 and IG9 sites are larger than those that came before them. This difference between the size and setting of each of the sampling sites and their respective zones of hydrological influence confer unique characteristics for each one.

## Database

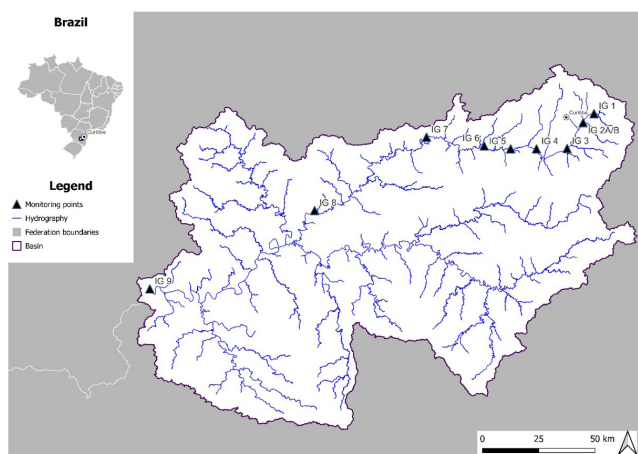
All the data used in this study was collected through the Project Integra II, a multi-institution, multidisciplinary water quality monitoring scientific effort performed on the Iguassu River, in southern Brazil, from 2005 through 2017. During this period, 415 surface water samples were collected in non-regular time-steps, from 10 unevenly spaced sampling sites along the first half of the course of the Iguassu River.

The dataset is composed of the observations of the concentrations of 11 traditional water quality parameters, as well as the concentrations of 4 contaminants of emerging concern. The parameters which were monitored were: biological oxygen demand (BOD), particulate and dissolved organic carbon (POC and DOC, respectively), ammonia (NH<sub>4</sub>), organic and total nitrogen (NO<sub>Org</sub> and TN, respectively), nitrate and nitrite (NO<sub>2</sub> and NO<sub>3</sub>, respectively), orthophosphate (PO<sub>4</sub>), total phosphorus (TP) and dissolved oxygen (DO). The 4 contaminants of emerging are: caffeine (CAF), estradiol (E1), ethinylestradiol (EE1) and estrone (E2). The latter was ultimately dropped from the analysis for lacking informational entropy, resulting in an analysis that encompassed only 3 contaminants of emerging concern.

The database has varying amounts of missing values for each of the parameters. Table 2, next, shows the number of observations, missing values, average and standard deviation for each of the studied parameters.

The missing values of POC, PO<sub>4</sub>, CAF, E1 and EE1 were mostly caused by alterations in the monitoring strategy during the study period. DOC monitoring started only in 2009, CAF and the hormones were measured starting in 2012, and POC has only been observed from 2012 to 2014. The other missing values were caused by machine malfunction and are spaced throughout the study period. Besides the considerable number of missing values within the dataset, the samples showed high variability, as their mean values for all parameters are smaller than their standard deviations.

Due to the lack of informational entropy within the estrone example dataset (all examples for this hormone's concentration



**Figure 1.** Map of the study area within the Iguassu River Basin.

**Table 2.** Status of the dataset and measures of central tendency and dispersion.

PAR	Missing Values	N. Obs	Mean	STD	PAR	Missing Values	N. Obs	Mean	STD
BOD	39	377	16.07	16.27	NT	34	382	9.50	11.10
POC	339	77	5.94	6.13	PT	9	407	0.79	1.38
DOC	44	372	7.90	5.37	PO <sub>4</sub>	155	261	0.47	1.30
NH <sub>4</sub>	23	393	5.96	8.92	OD	40	376	3.39	2.27
NO <sub>2</sub>	7	409	0.22	0.33	CAF	307	109	2.30	4.31
NO <sub>3</sub>	7	389	0.94	2.10	E1	306	110	0.17	0.46
NO <sub>Org</sub>	58	358	3.31	4.61	EE1	306	110	0.18	0.52

were zero or below the limit of quantification) the analysis of imputation of new data for this parameter was dropped, as any attempt to generate an artificial intelligence method would only allow for the replication of null results.

### Network architecture and data imputation process

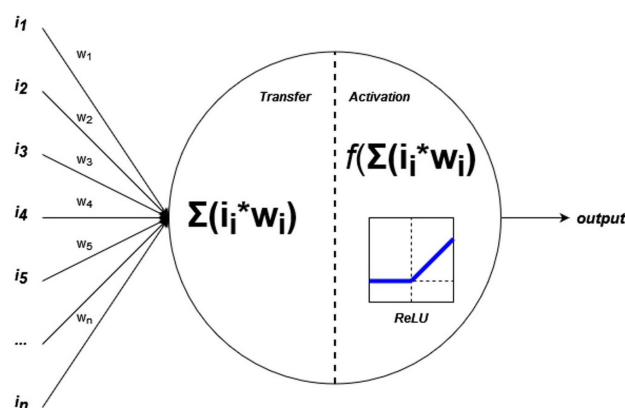
An artificial neural network (ANN) is a computational model which has the objective of understanding and predicting an intrinsic complex non-linear dependency relationship between a set of input variables (predictors) and a set of output variables (targets). This net is based on mathematical neurons, which are disposed into interconnected layers that allow computing high level models from simple neural models (Khalil et al., 2011). The system receives the input data and uses many parameters to compute the solution, which need to be estimated. So, a neural network is an iterative method that focuses on minimizing the error in prediction (the performance metric for this study was the MSE – mean squared error). This is achieved by exposing the model to data examples with previously known values for inputs and outputs, this process is called the training of the model (Jiang et al., 2013; Wang et al., 2019; Bansal & Geetha, 2020).

The input variables are inserted into the input layer of the network. Each node of this layer receives the input signals and computes a result. As each node is connected to each of the nodes of the next (now called hidden) layer, its output is passed to a subsequent hidden layer (feed forward) or, finally, the output layer. Each connection between neurons has a weight assigned to it, which will be updated, by an iterative process, to minimize the difference between the observed (real) and modeled results. This iterative correction is called back-propagation, as it spreads from the output nodes towards the input (Ruben et al., 2018; Mitrovic et al., 2019). This kind of net is called a feed forward artificial neural net (FFANN, as the input information has a directional path towards the output nodes), trained with the propagation algorithm, or multi-layer perceptronic neural network (MLPNN), the nomenclature is not consistent within the literature, therefore, for this study it will be used FFANN to identify this sort of ANN that is used here.

The operations that occur inside a hidden artificial neuron are usually arranged into two compartments. First, the Transfer Function section is responsible for summing the information of each weighted input. It then sends the value of this summation to the Activation Function compartment, which applies a user-defined function to modulate the output before passing it to another neuron within the net (Khalil et al., 2011). An illustration of an artificial neuron is presented on Figure 2. Note that in this case a ReLU (rectified linear unit) activation function is applied.

As previously stated, networks are arranged in interconnected layers of neurons. This allows representing and understanding more complex underlying patterns within a dataset.

In this study, it was verified if it would be possible to fill the missed values of a given water quality parameter, computing a value from other variables using a neural network. That means that it was evaluated the viability of predicting one of the variables using the others as information source, assuming that there is a certain dependency between the variables and that the neural network was able to detect this relationship. For this purpose,



**Figure 2.** Function of a generic neuron in a feed-forward neural network with a ReLU activation function.

it is important that enough samples with several parameters are available. For example, in this study case, for the 372 samples in which the DOC concentration was measured, 348 had also an associated measurement of BOD, while only 77 of these samples had an associated concentration for POC. Note that as there were only 110 observations for the contaminants of emerging concern, they were not considered as possible input parameters. By taking these values into consideration, the inputs that would be fed into each of the models are presented on Table 3.

To verify this hypothesis, neural networks were used varying the architecture. A script was written to measure the performance of different neural networks architectures using the same input data. The script trained 1641 different configurations for each of the 14 parameters, varying the number of nodes for the first, second, and third layers. The number of nodes for the first layer varied from 1 to 32, the second varied from 0 to 16, and the third varied from 0 to 8. When the value of layer reached zero within the iteration, it was taken away from the overall model, if the second layer became zero, so would the third. This preliminary sieve for the definition of the specific network architecture experienced 100 epochs.

After obtaining the performance of each configuration, the 5 settings which presented the best performance were re-trained with different epochs to circumvent possible under or overfitting caused by a non-optimal epoch number. The experiments started with 150 and finished with 1,000 epochs, with an iterative step of 50. The results for the network architectures which presented the best performance for each of the variables that would be used to generate the new imputed values are presented on Table 4.

These neural network models were trained by being exposed to examples which had all input data observed, not missing a single value within the input vector, as well as having all the input parameters 0-1 standardized. To determine the accuracy of the models, approximately 10% of the examples (aside from the 20% which were separate inside the algorithm to produce the testing accuracy values) were set aside for validation purposes. Table 5, next, presents the number of valid complete examples present on the database for each parameter, and the number of random examples which were excluded from the training dataset for validation.

**Table 3.** Set of input parameters used for training the models for each of the output parameters.

		INPUTS										
OUTPUT	PAR	BOD	COP	COD	NH4	NO2	NO3	NOrg	NT	PT	PO4	DO
	BOD			X	X	X	X		X	X		X
	POC	X		X	X	X	X		X	X	X	X
	DOC	X			X	X	X		X	X		X
	NH4	X		X					X	X		X
	NO2	X		X	X		X		X	X		X
	NO3	X		X	X	X			X	X		X
	NOrg	X		X	X	X	X		X	X		X
	NT	X		X	X	X	X			X		X
	PT	X		X	X	X	X		X			X
	PO4				X	X	X		X	X		X
	DO	X		X	X	X	X		X	X		
	CAF	X			X	X	X		X	X		X
	E1	X			X	X	X		X	X		X
	EE1	X			X	X	X		X	X		X

The parameters which would be used for validation were selected randomly for each different training run, because if these parameters were locked from the start, the models' measured accuracy would choose the ones which would better predict solely those 10% of examples previously set. The accuracy of each model's validation capability was measured through the root mean square error.

The overall imputation process was iterative. The chosen FFANN models were then applied to situations in which the output parameter was missing, but the input parameters were filled. Following stages would allow for the imputation of new values, as some input vectors which had missing values were completed by the previous imputation stage. Therefore, more than one imputation stage was performed for filling the dataset.

## RESULTS AND DISCUSSION

In the present study, the values for the concentrations of the biological oxygen demand (BOD), particulate and dissolved organic carbon (POC, DOC), ammonia nitrogen (NH<sub>4</sub>), nitrate (NO<sub>3</sub>), nitrite (NO<sub>2</sub>), organic nitrogen (NOrg), total nitrogen and phosphorus (TN, TP), orthophosphate (PO<sub>4</sub>), dissolved oxygen (DO), caffeine (CAF), estradiol (E1) and ethinylestradiol (EE1), on specific sampling dates in which collection of veritable data points were not performed, were predicted through artificial intelligence modeling.

The performance for each of the chosen models are presented on Table 6.

The overall performance of each of the developed models, when compared to those present on literature (Table 7), used as benchmarks, were acceptable, as the values for the validation errors did not differ greatly from previous studies, and were within the range of the lowest errors presented. But, as the RMSE values are intrinsically dependent on the scale of the measurements, the true extent of the performances of different models are difficult to assess without the application of the model on a standardized dataset.

Though the RMSE metric has been widely used for the analysis of the error of a predictive artificial intelligence model,

**Table 4.** Network architecture (number of nodes per layer) for the models of each of the output parameters.

PAR	1 <sup>st</sup> Layer	2 <sup>nd</sup> Layer	3 <sup>rd</sup> Layer
BOD	31	10	4
POC	12	15	5
DOC	12	6	8
NH <sub>4</sub>	25	8	3
N-Org	18	11	7
NO <sub>2</sub>	17	10	3
NO <sub>3</sub>	23	9	8
NT	24	10	-
PT	10	15	4
PO <sub>4</sub>	21	14	8
OD	10	11	4
CAF	20	10	4
E1	9	11	-
EE1	6	9	3

there are no set reference margins which allow for the general evaluation of a model - though several studies have been performed to estimate levels of goodness of fit for different environmental variables (e.g. the recommendation of the Wisconsin Department of Natural Resources Bureau (2007) which classifies models as Excellent if the error lies within a margin of 0.3 mg L<sup>-1</sup>, for the prediction of DO concentrations, such is the case for the proposed model), no consensus has been achieved (Kouadri et al., 2021; Boursalie et al., 2022). Therefore, the analysis of overall performance of a model is dependent on several factors, such as the scale of the input and output variable, the scope of the model/project, and the complexity of the system. For the scope of this study, due to the high variability of the input variables and the complexity of the system being represented, the errors presented by Table 6 (which vary from 0.04 mg L<sup>-1</sup> to 0.978 mg L<sup>-1</sup>) might be defined as a good representation of the starting dataset.

No studies have been performed to analyze the viability of predicting/modeling the concentrations of CAF, E1 or EE1 (Tiyasha et al., 2020).

**Table 5.** Number of valid examples and their separation in training and validation sets for each of the output parameters.

PAR	Valid examples	Train	Validation	PAR	Valid examples	Train	Validation
BOD	283	255	28	NT	283	255	28
POC	64	56	8	PT	283	255	28
DOC	283	255	28	PO4	238	214	24
NH4	283	255	28	OD	283	255	28
NO2	283	255	28	CAF	93	84	9
NO3	283	255	28	E1	93	84	9
NOrg	261	235	26	EE1	93	84	9

**Table 6.** Overall performance for each of the models implemented.

PAR	Train RMSE	Validation RMSE	Error STD	PAR	Train RMSE	Validation RMSE	Error STD
<b>BOD</b>	0.723	0.978	1.398	<b>TN</b>	0.217	0.446	0.663
<b>POC</b>	0.408	0.531	0.715	<b>TP</b>	0.034	0.096	0.139
<b>DOC</b>	0.222	0.423	0.729	<b>PO4</b>	0.017	0.054	0.114
<b>NH4</b>	0.183	0.285	0.633	<b>OD</b>	0.114	0.221	0.488
<b>NO2</b>	0.013	0.098	0.088	<b>CAF</b>	0.212	0.233	0.410
<b>NO3</b>	0.050	0.175	0.183	<b>E1</b>	0.040	0.044	0.059
<b>NOrg</b>	0.114	0.638	0.704	<b>EE1</b>	0.044	0.042	0.064

For the values obtained in this study, as the imputation method applied was iterative, the amount of predicted data for each parameter was different for each iteration, as shown in Table 8. None of the parameters had a new example opened up, by filling its input vector, by the third iteration.

Due to the different starting stages of each of the parameters input list, as well as the amount of missing data within each one, the final number of examples still had missing values, as no new vectors were opened up by the completion of new input vectors that would allow for new prediction of missing values. Table 9 presents the number of final values for each of the parameters.

The percentage of prediction between the different parameters were not equal. As expected, those parameters which presented the least amount of filled examples were the ones which had the largest number of imputed values. Therefore, the values for caffeine, estradiol and ethinylestradiol were the ones which presented the highest rate of filled in values. A total of 1.370 values were imputed overall, 61.57% of which are represented by the imputation of values for the concentrations of the contaminants of emerging concern.

The distribution for the imputed values for the contaminants of emerging concern are presented on Figure 3 (Caffeine), Figure 4 (estradiol) and Figure 5 (ethinylestradiol).

For the hormones, the peaks of the distribution of modeled values were consistently smaller than those measured, which might be explained by the presence of a high percentage of true zeros examples. Since E1 had 57.2% of its training examples as true zeros, and EE1 presented a true zero percentage of 54.5%, which contrasts to the low percentage of true zero caffeine examples of 17.3%. This could have overwhelmed the model to tend to lower concentrations, yet different behaviors shown by the input parameters during the period in which the contaminants of

**Table 7.** Performance of past studies which predicted the concentrations for the output parameters.

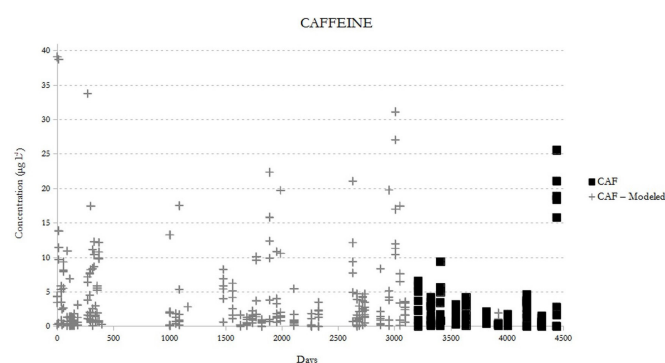
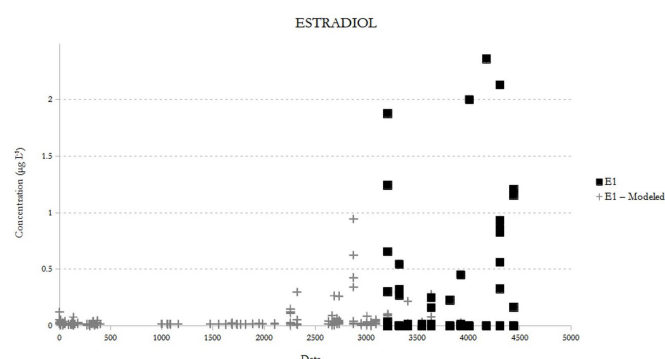
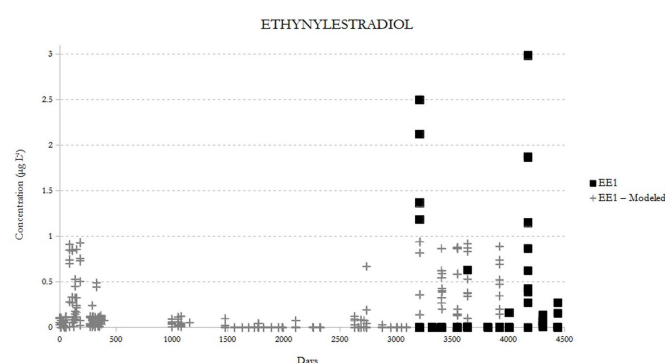
	Validation RMSE	Author
<b>BOD</b>	0.32	Ahmadi et al. (2018)
	1.67	Ahmed & Shah (2017)
	0.42	Banejad & Olyaie (2011)
<b>POC</b>	0.10	Buchard-Levine et al. (2014)
	4.43	Zhang et al. (2021)
<b>DOC</b>	0.20	Buchard-Levine et al. (2014)
	0.30	Zhou (2020)
<b>NH4</b>	2.05	Suen & Eheart (2003)
	0.33	Buchard-Levine et al. (2014)
<b>NOrg</b>	0.15	Zhou (2020)
<b>NO2</b>	0.10	Ha et al. (2020)
	0.66	Li et al. (2020a)
<b>NO3</b>	0.16	Ha et al. (2020)
	0.56	Li et al. (2020b)
<b>TN</b>	1.34	Shen et al. (2019)
	0.11	Lu et al. (2022)
<b>TP</b>	0.25	Shen et al. (2019)
	0.02	Lu et al. (2022)
	0.01	Ha et al. (2020)
<b>PO4</b>	0.01	Ha et al. (2020)
	0.47	Heddad (2016)
<b>DO</b>	0.73	Csábrági et al. (2019)
	0.88	Abba et al. (2017)
	0.84	Banejad & Olyaie (2011)

emerging concern were not measured could have led to a model representation that results in lower concentrations overall.

As another product of model generation, the underlying function between the output parameters and each of its input

**Table 8.** Number of filled values in each iteration for the output parameters.

PAR	1 <sup>st</sup> iter	2 <sup>nd</sup> iter	3 <sup>rd</sup> iter	PAR	1 <sup>st</sup> iter	2 <sup>nd</sup> iter	3 <sup>rd</sup> iter
BOD	20	-	-	TN	-	-	-
POC	128	135	34	TP	1	-	-
DOC	20	-	-	PO4	92	34	2
NH4	10	-	-	DO	32	-	-
NO2	-	-	-	CAF	145	101	39
NO3	3	-	-	E1	145	100	34
NOrg	22	7	-	EE1	145	100	34

**Figure 3.** Distribution of observed and modeled examples of caffeine through the study period.**Figure 4.** Distribution of observed and modeled examples of estradiol through the study period.**Figure 5.** Distribution of observed and modeled examples of ethinylestradiol through the study period.**Table 9.** Total number of filled values by output parameter.

PAR	Final number of examples	PAR	Final number of examples
BOD	397	TN	382
POC	374	TP	407
DOC	392	PO4	388
NH4	403	DO	408
NO2	409	CAF	391
NO3	392	E1	385
NOrg	387	EE1	385

parameters may be explored by fixing the values for all non-screening parameters and iterating the values for the target input. For example, for the relationship found between BOD and CAF, every other input had its value set to zero, and the model was asked to predict the probable value of caffeine (CAF) while BOD varied from 0 to 200 mg L<sup>-1</sup>. The limits of variation for the x-axis were determined as a rounding of the double of the highest observed concentration for each of the parameters used as input for each specific output (the contaminants of emerging concern). Figure 6 presents the behavior presented by each of the input variables for the contaminants of emerging concern.

The behavior of the relations between the input variables and their output varied for each of the contaminants of emerging concern. The predicted concentrations for caffeine and estradiol presented a linear rise in relation to the concentrations of biological oxygen demand and ammonia nitrogen, which were expected, as higher concentrations of these parameters have been previously linked to an increase in water quality degradation (Mizukawa et al., 2019). This behavior was the opposite of the one presented by the dissolved oxygen (for caffeine and ethinylestradiol), as the fall in concentration of this parameter is usually linked to the discharge of wastewater, for the modeling of the estradiol concentration it did not present any significant relationship, this could be explained by the sensibility of the parameter, as variations on this parameter depend not only on chemical and biological variations but also physical ones, as higher temperatures diminish the maximum saturation of oxygen on the water (Li et al., 2020b).

The relationship for some of the input variables and ethinylestradiol, BOD, NH<sub>4</sub>, and total phosphorus, presented an initial peak of 0.17 µg L<sup>-1</sup> for a null concentration of the parameters, followed by close to zero contribution as the concentration of said input parameter rose. This behavior is expected when the model cancels out an input variable, and could be understood as

the lack of informational entropy offered by the variable when modeling the output.

The last step for understanding the relationship between the variables used as input and the output was to establish the improvement (or decay) in performance when an input parameter was fixed to its mean for all examples. This operation allows for the determination of the relevance of the fixed parameter on the result produced by the entire model. As a parameter is fixed on its mean, the overall performance of the model is altered; the more the performance goes down, the more relevant that parameter was to the output. Therefore, it is possible to assume that the runs in which the performance was stable (or even increased) by the fixation of a parameter, that parameter is irrelevant to the overall result of the model.

Table 10, next, presents the results for the effects of fixing each input parameter onto each of the output models.

Each of the output models presented a different relationship between its variables. The organic matter measures (biochemical oxygen demand, particulate and dissolved organic carbon) did not present specific characteristics about the parameters in their inputs with two exceptions, both regarding DOC: The presence of ammonia nitrogen in the modeling was highly correlated to the performance of the model, while the removal of the total

phosphorus from the inputs produced results which were 29.4% more accurate.

For the nitrogenated parameters (ammonia nitrogen, organic nitrogen, nitrate, nitrite and total nitrogen) their inter-relationship was shown to be most relevant. Yet, while ammonia nitrogen and total nitrogen had their performance dependent of BOD, the modelling for organic nitrogen presented the highest improvement in performance while the BOD levels were fixed at its mean. The phosphorated compounds models were also dependent on the presence of total nitrogen within its input parameters.

For the contaminants of emerging concern, the distribution of biochemical oxygen demand, total nitrogen and dissolved oxygen were the ones which presented the highest relevance towards the performance of the three models. For the caffeine model, nitrate did not present such relevance as the one pointed out by the hormones' models. Also, the parameters which had no effects on the modelling were the distribution of ammonia nitrogen on the ethinylestradiol output, nitrite for the caffeine output and nitrate for the estradiol output.

The dynamics of micropollutants in aquatic environments are difficult to completely understand, due to these compounds' intrinsic characteristics and their relations to anthropic activities. Thus, the response sensibility and the synergy of one compound to

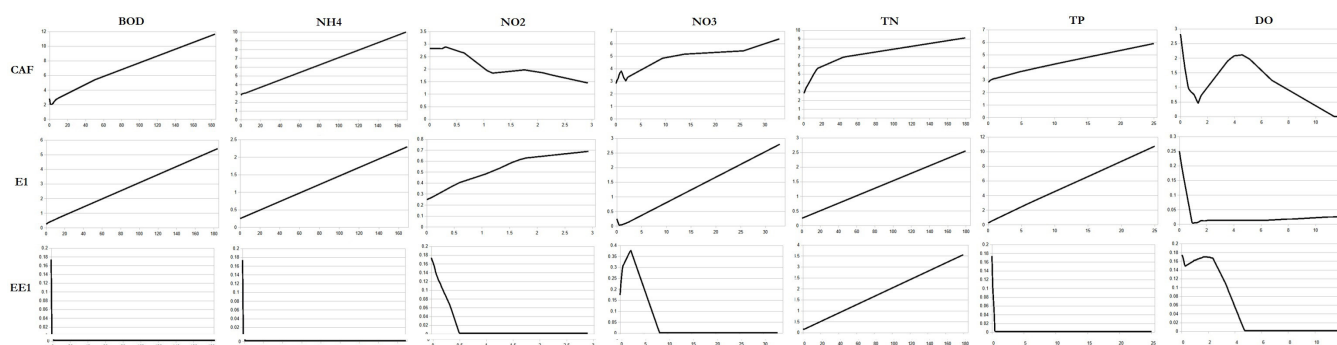


Figure 6. Behavior of the concentrations of the contaminants of emerging concern as each of their input behavior was the only varying aspect in the model.

Table 10. Performance for each of the output models when each input parameter is fixed in its mean value.

		FIXED PARAMETER								
		BOD	DOC	NH4	NO2	NO3	TN	TP	PO4	DO
OUTPUT PARAMETER	BOD	--	70.5	73.1	77.7	87	58.3	64.3	--	78.4
	POC	79.9	89.9	93.2	95.2	97.7	98.1	91.3	90.1	88.7
	DOC	64	--	36.9	93.6	74.2	55.6	129.4	--	64.3
	NH4	11.8	40.2	--	55.6	46.1	2.2	47.6	--	42.2
	Norg	132.3	99.2	38.3	61.2	41.8	1.2	95.8	--	122.4
	NO2	98.7	99.8	92.3	--	4.2	62.3	93.7	--	94.1
	NO3	66.5	85.6	70	7.3	--	18.3	108.3	--	83
	TN	18.1	110.4	2.8	22.6	14.8	--	10.8	--	39.8
	TP	23.1	102.4	9.8	76	82.6	11.9	--	--	67.6
	PO4	--	--	14.49	89.9	74.1	31.1	7.7	--	72.1
	DO	8.3	108.5	46.6	86.1	85.7	111.3	106.1	--	--
	CAF	56.6	--	47.3	87.2	97.36	28	51.1	--	77.1
	E1	61.3	--	65	97.8	51.4	52.8	84.9	--	53.3
EE1	54.3	--	100.1	99	57.1	55.7	97.6	--	63.1	



another is not completely known as well as their toxicological risks or the biological response to chronic exposure to these compounds, when exposed in a natural water setting (Routledge et al., 1998; Montagner et al., 2019; Santos et al., 2020). Yet, it is well established these compounds' great potential for endocrine disruption and deleterious health effects, especially on aquatic animals (Rocha et al., 2013; Luo et al., 2014). Aside from the compounds' inherent characteristics, there are environmental aspects that should be taken into account, such as the compounds' environmental half-life and recalcitrance, dilution caused by increased precipitation and other anthropic influxes, seasonal variance and changes in land cover and land use (Yoon et al., 2010; Wang et al., 2018; Yuan et al., 2020). As for the dilution and the seasonal variances, these might be the reason for the high rate of non-detection, or the concentrations being below the limit of quantification for the contaminants of emerging concern. Generally, higher concentrations of contaminants of emerging concern are linked to water quality degradation, which leads to the decrease in dissolved oxygen, and consequent increase of the concentrations of biochemical oxygen demand, ammonia nitrogen and phosphorus within the aquatic environment (Mizukawa et al., 2019).

The physical and chemical water quality parameters are indirect indications of anthropic activities which influence the water resources systems, while the concentrations of contaminants of emerging concern reflect a more accurate picture of the human impact on the aquatic environment, especially the influence caused by the discharge of wastewater. Therefore, it is necessary to consider the local and seasonal scenario which produced the changes in the physical and chemical parameters, as variations in the concentrations of OD, BOD, nitrogen and phosphorus may represent the natural configuration of the ecosystem being studied, or being a response to inefficient wastewater treatment (Yang et al., 2013; Berger et al., 2017). This is one of the great challenges of dealing with complex systems, which need to be mindfully studied. Therefore, advancements in technology and techniques to increase the quality of data collection and overall environmental monitoring are necessary to maintain up-to-date knowledge of the environment in which the researchers and water resource managers are included into.

## CONCLUSION

This study presented a modeling strategy to impute missing data of 11 traditional water quality parameters and 3 contaminants of emerging concern through the implementation of specific neural network models for each of the parameters analyzed. The results presented point to the viability of using feed-forward neural networks to predict missing data values in past sampling campaigns for the Iguassu River. One of the main concerns about the implementation of artificial intelligence models is their black-box characteristics, in which the input transformations within the model are not visible to the user, this problem was addressed by this study twice, aiming at exploring the intrinsic relationship between the input and output variables. Both methods of evaluating the importance of each variable in the prediction pointed to the chemical and physical validity of the relations explored by the models.

In this study, the results point to promising model performances for the data imputation of caffeine concentrations, using as input the concentrations of biochemical oxygen demand, ammonia nitrogen, total nitrogen and phosphorus and dissolved oxygen. The overall performance of model expected errors of a little more of 10% of the measured concentrations for this compound. Though this dataset was a product of one of the largest and most comprehensive water quality monitoring efforts in Brazil, it was still not possible to gather data with enough informational entropy to allow for a better imputation model for estradiol and ethinylestradiol (as the majority of the training examples were composed of true zeros). Also, as in this study the networks were trained on examples for all sampling sites, and each of the sites presents specific characteristics, other studies might benefit from performing a similar analysis on a per-site basis, or by grouping sites with similar characteristics.

The efforts to monitor and understand the aquatic environments are dependent on large monetary investments, as well as being susceptible to inconsistencies in sampling, which could possibly generate gaps in the dataset. Artificial intelligence methods have been shown to be a viable modeling alternative for understanding, and consequently predicting, complex patterns in high dimensional datasets, and the results obtained in this study point to the viability of modeling aquatic environments by feed-forward neural networks. Further studies should be performed to better understand the extent which the implementation of machine learning methods could improve the understanding of the environmental settings researchers, water resource managers and decision-makers are included in, as well as improve resource allocation in environmental monitoring efforts while potentially diminishing its financial burden.

## ACKNOWLEDGEMENTS

The authors thank the Coordination for the Improvement of Higher Education Personnel (CAPES) and the Council for Scientific and Technological Development (CNPq) for the funding. The authors also thank the Federal University of Parana (UFPR), Department of Hydraulics and Sanitation – Graduate Program in Environmental and Water Resources Engineering, and the Technological University of Parana (UTFPR), Academic Department for Chemistry and Biology.

## REFERENCES

- Abba, S. I., Hadi, S. J., & Abdullahi, J. (2017). River water modelling prediction using multi-linear regression, artificial neural network, and adaptive neuro-fuzzy inference system techniques. *Procedia Computer Science*, 120, 75-82. <http://dx.doi.org/10.1016/j.procs.2017.11.212>.
- Ahmadi, A., Fatemi, Z., & Nazari, S. (2018). Assessment of input data selection methods for BOD simulation using data-driven models: a case study. *Environmental Monitoring and Assessment*, 190(4), 239. <http://dx.doi.org/10.1007/s10661-018-6608-4>.
- Ahmed, A. A. M., & Shah, S. M. S. (2017). Application of adaptive neuro-fuzzy inference system (ANFIS) to estimate the biochemical

- oxygen demand (BOD) of Surma River. *Journal of King Saud University – Engineering Sciences*, 29(3), 237-243.
- Ahmed, A. N., Binti Othman, F., Abdulmohsin Afan, H., Khaleel Ibrahim, R., Ming Fai, C., Shabbir Hossain, M., Ehteram, M., & Elshafie, A. (2019). Machine learning methods for better water quality prediction. *Journal of Hydrology*, 578, 124084. <http://dx.doi.org/10.1016/j.jhydrol.2019.124084>
- Ahmed, M., Mumtaz, R., & Zaidi, M. H. (2021). Analysis of water quality indices and machine learning techniques for rating water pollution: a case study of Rawal Dam, Pakistan. *Water Science and Technology: Water Supply*, 21(6), 3225-3250. <http://dx.doi.org/10.2166/ws.2021.082>.
- Banejad, H. H., & Olyaei, E. H. (2011). Application of an artificial neural network model to rivers water quality indexes prediction: a case study. *The Journal of American Science*, 7(1), 60-65.
- Bansal, S., & Geetha, G. (2020). A machine learning approach towards automatic water quality monitoring. *Journal of Water Chemistry and Technology*, 42(5), 321-328. <http://dx.doi.org/10.3103/S1063455X20050045>.
- Berger, E., Haase, P., Kuemmerlen, M., Leps, M., Schäfer, R. B., & Sundermann, A. (2017). Water quality variables and pollution sources shaping stream macroinvertebrate communities. *The Science of the Total Environment*, 587-588, 1-10. <http://dx.doi.org/10.1016/j.scitotenv.2017.02.031>.
- Berrou, K., Roig, B., & Cadriere, A. (2021). Assessment of micropollutants toxicity by using a modified *Saccharomyces cerevisiae* model. *Environmental Pollution*, 291, 118-211. <https://doi.org/10.1016/j.envpol.2021.118211>.
- Boursalie, O., Samavi, R., & Doyle, T. E. (2022). Evaluation metrics for deep learning imputation models. *Studies in Computational Intelligence*, 13(10), 93-107. [http://dx.doi.org/10.1007/978-3-030-93080-6\\_22](http://dx.doi.org/10.1007/978-3-030-93080-6_22).
- Buchard-Levine, A., Liu, S., Vince, F., Li, M., & Ostfeld, A. (2014). A hybrid evolutionary data driven model for river water quality early warning. *Journal of Environmental Management*, 143(1), 8-16. <http://dx.doi.org/10.1016/j.jenvman.2014.04.017>.
- Csábrági, A., Molnár, S., Tanos, P., Kovács, J., Molnár, M., Szabó, I., & Hatvani, I. G. (2019). Estimation of dissolved oxygen in riverine ecosystems: comparison of differently optimized neural networks. *Ecological Engineering*, 138, 298-309. <http://dx.doi.org/10.1016/j.ecoleng.2019.07.023>.
- Galus, M., Kirischian, N., Higgins, S., Purdy, J., Chow, J., Ranganarajan, S., Li, H., Metcalfe, C., & Wilson, J. Y. (2013). Chronic, low concentration exposure to pharmaceuticals impacts multiple organ systems in zebrafish. *Aquatic Toxicology*, 132-133, 200-211. PMID:23375851.
- Giri, A. K. (1993). The genetic toxicology of paracetamol and aspirin: a review. *Mutation Research: Reviews in Genetic Toxicology*, 296(3), 199-210. PMID:7680103.
- Ha, N. T., Nguyen, H. Q., Truong, N. C. Q., Le, T. L., Thai, V. N., & Pham, T. L. (2020). Estimation of nitrogen and phosphorus concentrations from water quality surrogates using machine learning in the Tri An Reservoir, Vietnam. *Environmental Monitoring and Assessment*, 192(789), 789. <http://dx.doi.org/10.1007/s10661-020-08731-2>.
- Hayder, G., Kurniawan, I., & Mustafa, H. M. (2021). Implementation of machine learning methods for monitoring and predicting water quality parameters. *Biointerface Research in Applied Chemistry*, 11(2), 9285-9295.
- Heddam, S. (2016). Simultaneous modelling and forecasting of hourly dissolved oxygen concentration (DO) using radial basis function neural network (RBFNN) based approach: a case study from the Klamath River, Oregon, USA. *Modeling Earth Systems and Environment*, 2(135), 135. <http://dx.doi.org/10.1007/s40808-016-0197-4>.
- Isidori, M., Lavorgna, M., Nardelli, A., Parrella, A., Previtera, L., & Rubino, M. (2005). Ecotoxicity of naproxen and its phototransformation products. *The Science of the Total Environment*, 348(1-3), 93-101.
- Jiang, Y., Nan, Z., & Yang, S. (2013). Risk assessment of water quality using Monte Carlo simulation and artificial neural network method. *Journal of Environmental Management*, 122, 130-136. <http://dx.doi.org/10.1016/j.jenvman.2013.03.015>.
- Kamyab-Talesh, F., Mousavi, S. F., Khaledian, M., Yousefi-Falakdehi, O., & Norouzi-Masir, M. (2019). Prediction of water quality index by support vector machine: a case study in the Sefidrud Basin, Northern Iran. *Water Resources*, 46(1), 112-116. <http://dx.doi.org/10.1134/S0097807819010056>.
- Katipoglu-Yazan, T., Pala-Ozkok, I., Ubay-Cokgor, E., & Orhon, D. (2013). Acute impact of erythromycin and tetracycline on the kinetics of nitrification and organic carbon removal in mixed microbial culture. *Bioresource Technology*, 144, 410-419. <http://dx.doi.org/10.1016/j.biortech.2013.06.121>.
- Khalil, B., Ouarda, T. B. M. J., & St-Hilaire, A. (2011). Estimation of water quality characteristics at ungauged sites using artificial neural networks and canonical correlation analysis. *Journal of Hydrology*, 405(3-4), 277-287. <http://dx.doi.org/10.1016/j.jhydrol.2011.05.024>.
- Kidd, K. A., Blanchfield, P. J., Mills, K. H., Palace, V. P., Evans, R. E., Lazorchak, J. M., & Flick, R. W. (2007). Collapse of a fish population after exposure to a synthetic estrogen. *Proceedings of the National Academy of Sciences of the United States of America*, 104(21), 8897-8901. <http://dx.doi.org/10.1073/pnas.0609568104>.
- Kiesling, R. L., Elliott, S. M., Kammel, L. E., Choy, S. J., & Hummel, S. L. (2019). Predicting the occurrence of chemicals of emerging

- concern in surfacewater and sediment across the U.S. portion of the Great Lakes Basin. *The Science of the Total Environment*, 651, 838-850. <http://dx.doi.org/10.1016/j.scitotenv.2018.09.201>.
- Kouadri, S., Ebeltagi, A., Islam, A. R. T., & Kateb, S. (2021). Performance of machine learning methods in predicting water quality index based on irregular data set: application on Illizi region (Algerian southeast). *Applied Water Science*, 11(12), 190. <http://dx.doi.org/10.1007/s13201-021-01528-9>.
- Krishnaraj, A., & Deka, P. C. (2020). Spatial and temporal variations in river water quality of the Middle Ganga Basin using unsupervised machine learning techniques. *Environmental Monitoring and Assessment*, 192, 744-762.
- Li, S., Bhattari, R., Cooke, R. A., Verma, S., Huang, X., Markus, M., & Christianson, L. (2020a). Relative performance of different data mining techniques for nitrate concentration and load estimation in different type of watersheds. *Environmental Pollution*, 263, 114618.
- Li, W., Fang, H., Qin, G., Tan, X., Huang, Z., Zeng, F., Du, H., & Li, S. (2020b). Concentration estimation of dissolved oxygen in Pearl River Basin using input variable selection and machine learning techniques. *The Science of the Total Environment*, 731, 139099. <http://dx.doi.org/10.1016/j.scitotenv.2020.139099>.
- Liu, M., & Lu, J. (2014). Support vector machine: an alternative to artificial neuron network for water quality forecasting in an agricultural nonpoint source polluted river? *Environmental Science and Pollution Research International*, 21(18), 11036-11053.
- Lu, H., Yang, L., Fan, Y., Qian, X., & Liu, T. (2022). Novel simulation of aqueous total nitrogen and phosphorus concentrations in Taihu Lake with machine learning. *Environmental Research*, 204, 111940. <http://dx.doi.org/10.1016/j.envres.2021.111940>.
- Luo, Y., Guo, W., Ngo, H. H., Nghiem, L. D., Hai, F. I., Zhang, J., Liang, D., & Wang, X. C. (2014). A review on the occurrence of micropollutants in the aquatic environment and their fate and removal during wastewater treatment. *The Science of the Total Environment*, 473-474, 619-641. <http://dx.doi.org/10.1016/j.scitotenv.2013.12.065>.
- Mitrovic, T., Antanasijevic, D., Lazovic, S., Peric-Grujic, A., & Ristic, M. (2019). Virtual water quality monitoring at inactive monitoring sites using Monte Carlo optimized artificial neural networks: A case study of Danube River (Serbia). *The Science of the Total Environment*, 654, 1000-1009. <http://dx.doi.org/10.1016/j.scitotenv.2018.11.189>.
- Mizukawa, A., Fillipi, T. C., Peixoto, L. O. M., Scipioni, B., Leonardi, I. R., & Azevedo, J. C. R. (2019). Caffeine as a chemical tracer for contamination of urban rivers. *Revista Brasileira de Recursos Hídricos*, 24, e29.
- Montagner, C. C., Sodré, F. F., Acayaba, R. D., Vidal, C., Campestrini, I., Locatelli, M. A., Pescara, I. C., Albuquerque, A. F., Umbuzeiro, G. A., & Jardim, W. F. (2019). Ten years-snapshot of the occurrence of emerging contaminants in drinking: surface and ground waters and wastewaters from São Paulo State, Brazil. *Journal of the Brazilian Chemical Society*, 30(3), 614-632.
- Muthukrishnan, N., Maleki, F., Ovens, K., Reinhold, C., Forghani, B., & Forghani, R. (2020). Brief History of Artificial Intelligence. *Neuroimaging Clinics of North America*, 30(4), 393-399. <http://dx.doi.org/10.1016/j.nic.2020.07.004>.
- Oaks, J. L., Gilbert, M., Virani, M. Z., Watson, R. T., Meteyer, C. U., Rideout, B. A., Shivaprasad, H. L., Ahmed, S., Chaudry, M. J. I., Arshad, M., Mahmood, S., Ali, A., & Khan, A. A. (2004). Diclofenac residues as the cause of vulture population decline in Pakistan. *Nature*, 427(6975), 630-633. Retrieved in 2023, April 11, from <http://www.nature.com/articles/nature02317>
- Ooi, K. S., Chen, Z., Poh, P. E., & Cui, J. (2022). BOD5 prediction using machine learning methods. *Water Science and Technology: Water Supply*, 22(1), 1168-1183. <http://dx.doi.org/10.2166/ws.2021.202>.
- Park, Y., Lee, H. K., Shin, J. K., Chon, K., Kim, S., Cho, K. H., Kim, J. K., & Baek, S. S. (2021). A machine learning approach for early warning of cyanobacterial bloom outbreaks in a freshwater reservoir. *Journal of Environmental Management*, 255, 112415. <https://doi.org/10.1016/j.jenvman.2021.112415>.
- Rocha, M. J., Cruzeiro, C., & Rocha, E. (2013). Quantification of 17 endocrine disruptor compounds and their spatial and seasonal distribution in the Iberian Ave River and its coastline. *Toxicological and Environmental Chemistry*, 95(3), 386-399. <http://dx.doi.org/10.1080/02772248.2013.773002>.
- Routledge, E. J., Sheahan, D., Desbrow, C., Brighty, G. C., Waldock, M., & Sumpter, J. P. (1998). Identification of estrogenic chemicals in STW effluent. 2. In vivo responses in trout and roach. *Environmental Science & Technology*, 32(11), 1559-1565. <http://dx.doi.org/10.1021/es970796a>.
- Ruben, G. B., Zhang, K., Bao, H., & Ma, X. (2018). Application and sensitivity analysis of artificial neural network for prediction of chemical oxygen demand. *Water Resources Management*, 32(1), 273-283. <http://dx.doi.org/10.1007/s11269-017-1809-0>.
- Santos, A. V., Couto, C. F., Lebron, Y. A., Moreira, V. R., Foureaux, A. F. S., Reis, E. O., & Lange, L. C. (2020). Occurrence and risk assessment of pharmaceutically active compounds in water supply systems in Brazil. *The Science of the Total Environment*, 746, 141011. <http://dx.doi.org/10.1016/j.scitotenv.2020.141011>.
- Shen, J., Qin, Q., Wang, Y., & Sisson, M. (2019). A data-driven modeling approach for simulating algal blooms in the tidal freshwater of James River in response to riverine nutrient loading. *Ecological Modelling*, 398, 44-54. <http://dx.doi.org/10.1016/j.ecolmodel.2019.02.005>.
- Suen, J. P., & Eheart, W. (2003). Evaluation of neural networks for modelling nitrate concentration in rivers. *Journal of Water*

- Resources Planning and Management*, 129(6), 505-510. [http://dx.doi.org/10.1061/\(ASCE\)0733-9496\(2003\)129:6\(505\)](http://dx.doi.org/10.1061/(ASCE)0733-9496(2003)129:6(505)).
- Tiyasha, Tung, T. M., & Yaseen, Z. M. (2020). A survey on water quality modelling using artificial intelligence methods: 2000-2020. *Journal of Hydrology*, 585, 124670. <http://dx.doi.org/10.1016/j.jhydrol.2020.124670>.
- Wang, F., Wang, Y., Zhang, K., Hu, M., Weng, Q., & Zhang, H. (2021). Spatial heterogeneity modeling of water quality based on random forest regression and model interpretation. *Environmental Research*, 202, 111660. <http://dx.doi.org/10.1016/j.envres.2021.111660>.
- Wang, P., Yao, J., Wang, G., Hao, F., Shrestha, S., Xue, B., Xie, G., & Peng, Y. (2019). Exploring the application of artificial intelligence technology for identification of water pollution characteristics and tracing the source of water quality pollutants. *The Science of the Total Environment*, 693, 133440. <http://dx.doi.org/10.1016/j.scitotenv.2019.07.246>.
- Wang, S., Zhu, Z., He, J., Yue, X., Pan, J., & Wang, Z. (2018). Steroidal and phenolic endocrine disrupting chemicals (EDCs) in surface water of Bahe River, China: Distribution, bioaccumulation, risk assessment and estrogenic effect on *Hemiculter leucisculus*. *Environmental Pollution*, 243, 103-114. <http://dx.doi.org/10.1016/j.envpol.2018.08.063>.
- Wang, Y., Zheng, T., Zhao, Y., Jiang, J., Wang, Y., Guo, L., & Wang, P. (2013). Monthly water quality forecasting and uncertainty assessment via bootstrapped wavelet neural networks under missing data for Harbin, China. *Environmental Science and Pollution Research International*, 20(12), 8909-8923. <http://dx.doi.org/10.1007/s11356-013-1874-8>.
- Wisconsin Department of Natural Resources Bureau . (2007). *Guidelines for deployment of continuous devices, watershed management* (42 p.). Madison. Retrieved in 2022, December 18, from <https://dnr.wi.gov/water/wsSWIMSDocument.ashx?documentSeqNo=20724754>
- Woodhouse, P., & Muller, M. (2017). Water governance: an historical perspective on concurrent debates. *World Development*, 92, 225-241. <http://dx.doi.org/10.1016/j.worlddev.2016.11.014>.
- Yang, X., Chen, F., Meng, F., Xie, Y., Chen, H., Young, K., & Fu, W. (2013). Occurrence and fate of PPCPs and correlations with water quality parameters in urban riverine waters of the Pearl River Delta, South China. *Environmental Science and Pollution Research International*, 20(8), 5864-5875. <http://dx.doi.org/10.1007/s11356-013-1641-x>.
- Yoon, Y., Ryu, J., Oh, J., Choi, B.-G., & Snyder, S. A. (2010). Occurrence of endocrine disrupting compounds, pharmaceuticals, and personal care products in the Han River (Seoul, South Korea). *The Science of the Total Environment*, 408(3), 636-643. <http://dx.doi.org/10.1016/j.scitotenv.2009.10.049>.
- Yuan, X., Hu, J., Li, S., & Yu, M. (2020). Occurrence, fate, and mass balance of selected pharmaceutical and personal care products (PPCPs) in an urbanized river. *Environmental Pollution*, 266, 115340. <http://dx.doi.org/10.1016/j.envpol.2020.115340>.
- Zhang, Y., Yao, X., Wu, Q., Huang, Y., Zhou, Z., Yang, J., & Liu, X. (2021). Turbidity prediction of lake-type raw water using random forest model based on meteorological data: a case study of Tai lake, China. *Journal of Environmental Management*, 290, 112657. <http://dx.doi.org/10.1016/j.jenvman.2021.112657>.
- Zhou, Y. (2020). Real-time probabilistic forecasting of river water quality under data missing situation: deep learning plus post-processing techniques. *Journal of Hydrology*, 589, 125164. <http://dx.doi.org/10.1016/j.jhydrol.2020.125164>.

### Authors contributions

Luis Otávio Miranda Peixoto: Conceptualization, methodology, data curation, formal analysis, investigation, writing – original draft.

Bárbara Alves de Lima: Data analysis, writing – review & editing.

Camila de Carvalho Almeida: Data analysis, writing – review & editing.

Cristóvão Vicente Scapulatempo Fernandes: Data collection, supervision, writing – review & editing.

Jorge Antonio Silva Centeno: Data analysis, supervision, writing – review & editing.

Júlio César Rodrigues de Azevedo: Data collection, data analysis, funding acquisition, supervision, writing – review & editing.

**Editor-in-Chief:** Adilson Pinheiro

**Associated Editor:** Carlos Henrique Ribeiro Lima