Science of the Total Environment 927 (2024) 172340

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Enhancing water quality prediction for fluctuating missing data scenarios: A dynamic Bayesian network-based processing system to monitor cyanobacteria proliferation

M. Pazo[a,*], S. Gerassis[a], M. Araújo[a], I. Margarida Antunes[b], X. Rigueira[a]

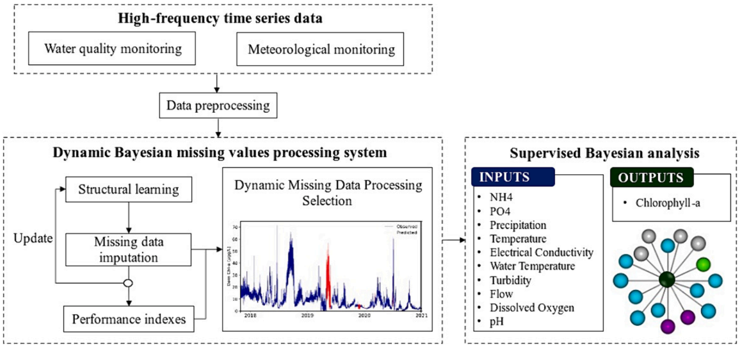
[a] *CINTECX, Universidade de Vigo, Grupo de Xestión Segura e Sostible de Recursos Minerais, Dpto. De Enxeñaría dos Recursos Naturais e Medio Ambiente, 36310 Vigo, Spain*

[b] *Institute of Earth Sciences (ICT), Pole of University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal*

HIGHLIGHTS

- Introduced dynamic Bayesian approach for water quality prediction.
- Compared DI, EBDI, and SEM dynamic imputation mehods.
- Identified SEM as the most effective method.
- BayesianML analysis identifies causes of cyanobacteria growth.

GRAPHICAL ABSTRACT



ARTICLE INFO

Editor: JV Cruz

Keywords:
Dynamic approach
Missing values
Bayesian learning
Cyanobacteria blooms
Livestock waste
Spain

ABSTRACT

Tackling the impact of missing data in water management is crucial to ensure the reliability of scientific research that informs decision-making processes in public health. The goal of this study is to ascertain the root causes associated with cyanobacteria proliferation under major missing data scenarios. For this purpose, a dynamic missing data management methodology is proposed using Bayesian Machine Learning for accurate surface water quality prediction of a river from Limia basin (Spain). The methodology used entails a sequence of analytical steps, starting with data pre-processing, followed by the selection of a reliable dynamic Bayesian missing value prediction system, leading finally to a supervised analysis of the behavioral patterns exhibited by cyanobacteria. For that, a total of 2,118,844 data points were used, with 205,316 (9.69 %) missing values identified. The machine learning testing showed the iterative structural expectation maximization (SEM) as the best performing algorithm, above the dynamic imputation (DI) and entropy-based dynamic imputation methods (EBDI), enhancing in some cases the accuracy of imputations by approximately 50 % in R2, RMSE, NRMSE, and logarithmic loss values. These findings can impact how data on water quality is being processed and studied, thus, opening the door for more reliable water management strategies that better inform public health decisions.

* Corresponding author.
E-mail address: maria.pazo@uvigo.gal (M. Pazo).

https://doi.org/10.1016/j.scitotenv.2024.172340
Received 25 January 2024; Received in revised form 5 April 2024; Accepted 7 April 2024
Available online 10 April 2024

1. Introduction

Water infrastructures and policies have been developed during centuries to ensure the agricultural, industrial, and urban development of the territory. Human development and access to water go hand in hand. However, the lack of adequate policy coordination and underinvestment have led to several water infrastructure and quality challenges in many parts of the world. Moreover, insufficient consideration has been given to the negative environmental implications arising from pollution sources and excessive depletion of natural resources (Xiao et al., 2021; Vigiak et al., 2023).

According to the United Nations Economic Commission for Europe (UNECE) and the World Health Organization (WHO), the projected area of the European Union with high water stress is estimated to increase from 19 %, in 2007, to 35 %, in 2070, affecting an additional 7 % of the population (Economic and Social Council, 2022). This scenario could lead to a situation in which water supplies would be less reliable and people's exposure to pathogens and harmful chemicals would increase (Pierrat et al., 2023). Factors like climate change, alongside global population growth, are intensifying the stress on our planet's water resources. Water overexploitation and inadequate management further exacerbate this crisis (Asif et al., 2023). For instance, as sea levels rise, coastal freshwater supplies face increasing saltwater intrusion, altering ecosystems, and threatening agriculture (Karimidastenaei et al., 2022). Additionally, the ever-growing demand for freshwater – essential for food production, which consumes nearly 90 % of the global supply (Scanlon et al., 2007) – is heightened by the widespread use of fertilizers, leading to increased waste and contamination (Mishra, 2023). These changes portray a stark picture of the challenges faced by communities worldwide, from small coastal towns to large agricultural regions, underscoring the urgent need for sustainable water management solutions. Over the past twenty years, there has been growing interest in the creation and application of hydrological models at global, continental, and national levels (Paul et al., 2021). At present, hydrological information is being increasingly exploited, demonstrating its great potential in the mitigation of adverse events, such as floods (e.g. Al-Sabhan et al., 2003; Chen et al., 2019), and in the development of new solutions, including adequate irrigation systems (Roy et al., 2023) or climate change adaptation strategies for water management (Abdelkarim et al., 2023; Muzammil et al., 2023; Özerol et al., 2020). Recent advances in water resource management demonstrate a notable shift towards employing emerging technologies. Convolutional neural networks (CNNs) have been applied by Sadeghi et al. (2020) for the integration of geographic and infrared data, enabling real-time precipitation monitoring. Gerassis et al. (2021) utilized Bayesian AutoML for an environmental impact assessment, presenting scenarios to understand basin contribution changes. Huang et al. (2022) integrate machine learning and deep learning algorithms to improve prediction and management strategies. Dhaoui et al. (2022) assess groundwater quality for irrigation using fuzzy logic in their research. These contributions underscore the movement towards digitalization and the use of decision-making models for informed and sustainable water management.. However, effectively harnessing the information provided by water infrastructures with quality monitoring systems remains a major challenge for researchers (Ezzati et al., 2023; Simpson et al., 2023). For example, the analysis of water quality time series data often presents difficulties due to non-normal distributions, seasonality, or runoff (e.g. Acock, 2005; Desbureaux et al., 2022; Albers, 2023). Furthermore, the quality of statistical analysis can be significantly influenced by the amount of missing data (Dong and Peng, 2013; Rigueira et al., 2023). Processing datasets with missing values is a frequent problem for which traditional methods (Ngouna et al., 2020) are of limited value, such as complete case deletion, pairwise deletion, or mean substitution (Lin, 2010; van Buuren, 2016).

Understanding the significance of accurate water quality assessment is crucial, especially when considering the challenges posed by incomplete datasets. Incomplete datasets in water quality studies can be a frequent situation (Tiyasha et al., 2020). They arise due to multiple factors, such as failures in the data-gathering sensors or environmental factors that hinder data collection. These gaps in data pose a substantial challenge in accurately assessing water quality, as they can lead to skewed interpretations and unreliable conclusions. Addressing the issue of missing values in water quality datasets is therefore crucial for the successful development and implementation of advanced water management solutions (Johnson et al., 2021). This necessarily requires a shift towards more advanced data handling techniques and analytical tools, which can not only manage the intricacies of water data but also enhance the applicability of the models developed. Based on that, an important question emerges: How can we address the challenge of studying water quality variables with significant amounts of missing values in our data samples?

To address this issue, this study proposes the introduction of Bayesian machine learning to facilitate the simulation of posterior data quality parameters distributions. The Bayesian model developed in this study aims to capture different types of data (e.g., physicochemical, hydraulic parameters or temporal variables) to integrate available hydrological and water quality information, as well as identifying the influence of the different anthropogenic activities and climate change. The obtained results will allow to the definition of new tailored probabilistic safety ranges for water quality, which could be translated into the design of scientifically, backed-up water quality management measures.

The possibility of modeling and inferring the aquatic system behavior and response mechanisms already represents an important advance for the decision-making and selection of the best available corrective measures (Zakwan et al., 2022), thereby facilitating public health and compliance with environmental objectives (Wang et al., 2019; European Commission, 2022). Notwithstanding this, the central goal of this investigation is to identify, analyze and characterize the significance of dynamic missing data management using Bayesian networks for accurate water quality prediction. This methodology is applied to a stream from Limia river basin in the autonomous region of Galicia (Spain), where missing data is a major problem that impedes to ascertain the root causes associated to high cyanobacteria occurrence. The excessive application of nitrates, pesticides, and other agrochemicals associated to the agricultural sector is typically the main degradation factor in in-land and coastal waters (e.g. Evans et al., 2019; Wiering et al., 2020). One of the most worrying problems is cyanobacteria and microalgae blooms whose appearance can prejudice water reservoirs and other aquatic ecosystems that provide valuable services to fishing, wildlife, recreation, and drinking water with a crucial ecological importance (Ho et al., 2021; Acuña-Alonso et al., 2021). These blooms often produce potent toxins, and their harshness and impact worsen when factors such as excess nutrients and climate change are combined (O'Neil et al., 2012; Boelee et al., 2019).

To shed light on that challenge, this study aims also to benefit from the introduction of innovative AI-based algorithms embedded in Bayesian machine learning processes. This approach in combination with the expert knowledge gathered over years allows to unveil and validate hidden relationships between the hydrological and water quality parameters. Previous studies have evaluated this Limia ecosystem negatively affected by the presence of agricultural and livestock activities, trying to understand the influence of hydrological or physical-chemical factors on the proliferation of cyanobacteria (Carballeira et al., 2018; Acuña-Alonso et al., 2020; Garzon-Vidueira et al., 2020; Moron-Lopez et al., 2021). However, the behavior patterns of the reservoir are not well known, and the problem persists.

2. Materials and methods

2.1. Study area and data collection

The data source was collected from three ecologically sensitive

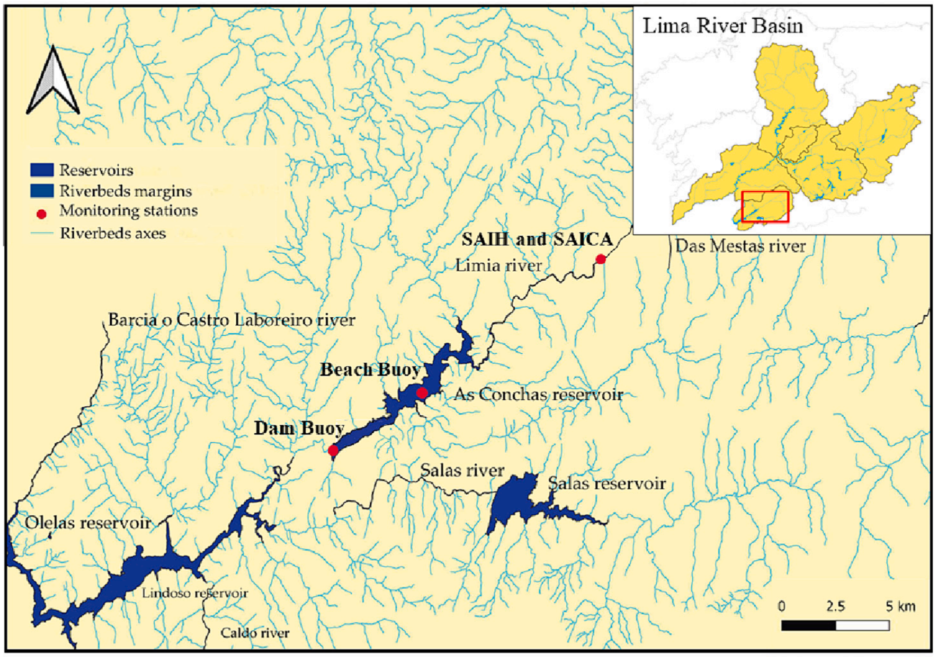


Fig. 1. Hydrography of the Limia River up to the As Conchas reservoir. Monitoring and control networks SAIH, SAICA, Chl-a dam buoy, and Chl-a beach buoy at three points of special sensitivity.

Table 1
Descriptive statistics for Chl-a concentration, water quality parameters, and meteorological data from a 3-year period (Oct 2017–Jan 2021).

Parameter	Unit	Mean ± Sd	Min	Max	Missing values (n)
Dam Chl-a	μg/L	9.10 ± 8.57	–	133.87	3.84 % (4280)
Beach Chl-a	μg/L	8.87 ± 7.98	–	135.84	1.47 % (1636)
NH_4	mg/L	0.07 ± 0.07	–	0.38	32.83 % (36624)
PO_4^{3-}	mg/L	0.08 ± 0.15	–	10.86	8.69 % (9695)
EC	μS/cm	102.25 ± 25.17	27.00	206.00	42.69 % (47616)
DO	mg/L	6.90 ± 2.79	0.50	12.80	24.86 % (27728)
pH	u. pH	6.49 ± 0.32	5.10	7.80	21.06 % (23495)
Water temperature	°C	12.38 ± 4.65	1.80	24.10	24.89 % (27762)
Turbidity	NTU	5.73 ± 10.78	0.00	200.00	22.36 % (24947)
Flow	m^3/s	5.95 ± 9.62	0.12	121.12	0.01 % (12)
Water level	m	0.26 ± 0.23	0.03	1.97	0.01 % (12)
Precipitation	mm	0.029 ± 0.17	–	13.60	1.34 % (1497)
Temperature	°C	11.80 ± 7.48	−8.30	39.40	0.01 % (12)

– Not detected. Total missing values: 205,316.

locations in the Limia riverbed, covering the stretch from the riverbed to the As Conchas reservoir in Galicia, Spain (Fig. 1). The data collection period spanned from October 2017 to January 2021, with water measurements taken every 15 min (2,118,844 data points in total).

The data series obtained could be classified into four main categories: (1) hydrological characteristics; (2) physicochemical parameters; (3) chlorophyll-a concentration; and (4) seasonality. More concretely, physicochemical, and hydraulic parameters were recorded using two different hydrographic systems deployed in Spain. Firstly, the Automatic Hydrological Information System (SAIH) for real-time monitoring of water. Secondly, the Automatic Water Quality Information System (SAICA) for water quality monitoring. The variables recorded with SAIH were river flow (m^3/s), air temperature (°C) and precipitation (mm), whereas for SAICA were dissolved oxygen (DO; mg/L), pH (u.pH), turbidity (NTU), electrical conductivity (EC; μS/cm), water temperature (°C), water level (m), ammonium (NH_4; mg/L) and phosphate (PO_4; mg/L) contents.

Additionally, high-frequency chlorophyll-a (Chl-a) concentration was extracted from a publicly accessible repository (Mozo et al., 2022). The Chl-a concentration was measured using two buoys anchored at specific locations in the central region of the reservoir, with an approximate spatial distance of 4 km, as reported by Moron-Lopez et al. (2021) and Mozo et al. (2022).

In the analyzed dataset, it was identified that 9.69 % (205,316 measures) of the data points are missing. This absence of data exhibits a considerable variation across different variables. Specifically, the missing data percentage ranges from a minimal 0.01 % in variables like river flow or air temperature, to a high of 42.69 % in variables, such as EC.

The descriptive statistics of the collected data were calculated and include the mean, standard deviation (sd), minimum value (Min), maximum value (Max), as well as the percentage and number of missing values for each variable (Table 1).

2.2. Bayesian machine learning

Bayesian networks are directed acyclic graphs (DAG) (G) created to evaluate causal and probabilistic relationships between variables, where the nodes of these models represent the domain variables, and the linking arcs represent the direct dependence relationships (Pearl, 1988). Consequently, G leads to the factorization (Scutari, 2019):

$$P(X|G,\theta) = \prod_{i=1}^{N} P(X_i|\Pi_{X_i},\theta_{X_i}), \tag{1}$$

where the joint probability distribution (JPD) of a random set of variables $X = \{X_1, \dots, X_n\}$, with parameters θ, splits into individual local distributions for each X_i (with parameters $\theta_{X_i}, \left. \bigcup_{X_i \epsilon X} \theta_{X_i} = \theta \right)$ dependent on its parent variables Π_{X_i}. Furthermore, these local distributions are mathematically calculated based on (Heckerman et al., 1995):

Table 2
Evaluation of unsupervised structural learning algorithms: MDL values and computational times.

Algorithms	DI		EBDI		SEM	
	MDL	Time[a]	MDL	Time[a]	MDL	Time[a]
MWST	3,236,262.98	1 m 34 s	3,164,965.88	1 m 1 s	3,150,576.43	15 m 15 s
Taboo	3,195,219.59	14 m 37 s	3,153,398.79	1 m 44 s	3,122,140.33	21 m 46 s
EQ	3,149,413.93	23 m 19 s	3,185,068.90	6 m 42 s	3,154,098.36	32 m 30 s
TabooEQ	3,179,329.98	5 h 56 m	3,172,223.74	2 h 16 m	3,154,964.49	4 h 26 m
SopLEQ	3,145,272.40	56 m 16 s	3,166,142.03	7 m 36 s	3,115,917.14	33 m 49 s

[a] The computational analyses were conducted on an Intel Core i9-12700F, Z690 motherboard, 64 GB DDR5 RAM, 1 TB NVMe PCIe SSD, and NVIDIA RTX 3070 8GB GPU.

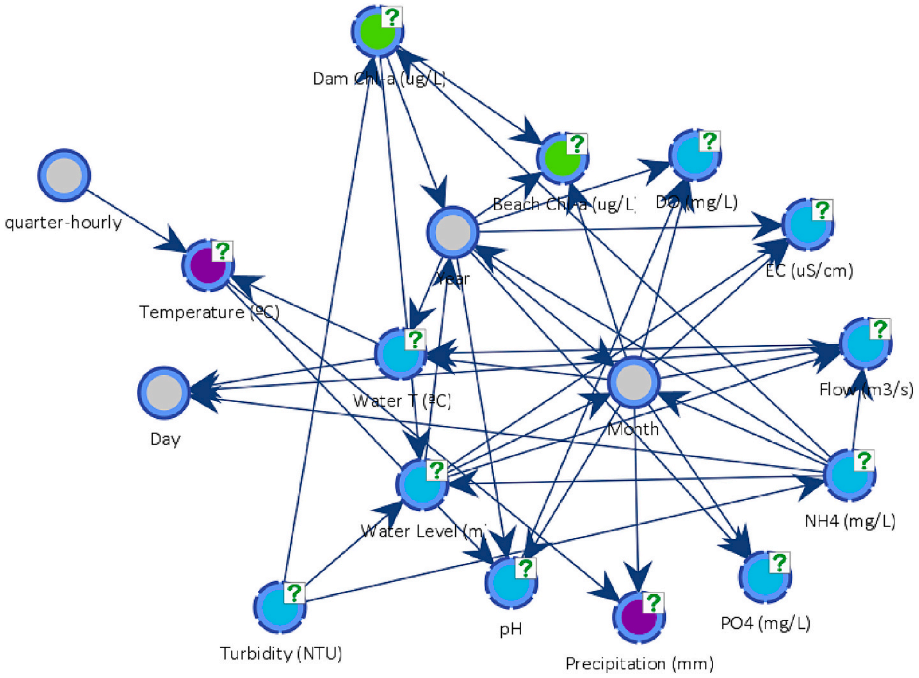


Fig. 2. Radial Layout of SEM Unsupervised Bayesian Network Using the SopLEQ Algorithm. Node colors classify variable types: green for cyanobacteria, blue for quality indicators, violet for meteorological, and gray for temporal variables. Nodes with missing values are marked with a question mark.

Table 3
Performance comparison of DI, EBBI and SEM models for water quality parameter prediction.

Imputations	R^2	RMSE	NRMSE	Precision	Log-loss	Calibration
DI	0.69	3.72	4.99 %	71.69 %	0.27	91.79 %
EBDI	0.71	4.04	5.37 %	79.04 %	0.26	92.74 %
SEM	0.76	2.92	3.85 %	80.00 %	0.22	92.79 %

$$X_i|\Pi_{X_i} \sim \text{Mul}\left(\pi_{ik|j}\right), \pi_{ik|j} = P(X_i = k|\Pi_{X_i} = j); \quad (2)$$

where parameters $\pi_{ik|j}$ correspond to the conditional probabilities of X_i, considering the distinct configurations of their parent values.

In this research, Bayesian networks, based on Judea Pearl's probabilistic theory, were employed to model a complex system. Their application in this study was twofold: firstly, to facilitate accurate predictions, and secondly, to elucidate causal and dependency relationships among variables. This dual capability is essential in machine learning for managing uncertainty, optimizing hyperparameters, and adapting to different data scenarios. Consequently, Bayesian networks were key in decision-making processes, especially under conditions of uncertainty associated to missing data.

2.3. Missing data mechanisms

Regarding the mechanisms underlying missing data, several researchers (e.g. Rubin, 1976; Little and Rubin, 2002; Little and Rubin, 2019) conducted extensive analysis and provided a widely accepted understanding of the nature and implications of missing data (Lin and Tsai, 2020; Ijadi Maghsoodi et al., 2023). Three missingness mechanisms are recognized: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR) (Rubin, 1976). MCAR, occurs when the absence of data is independent of both observed and unobserved variables, making the available data a representative sample of the total dataset. In contrast, MAR refers to situations where the missing data can be explained by observed data. Lastly, MNAR occurs when the missingness is a function of the not observed data.

2.4. Dynamic processing of missing values

In this study, dynamic inference techniques allow leveraging the advantages of structural algorithms, such as the ability to handle large complex datasets and to capture non-linear relationships. Similarly, while previous research has examined the trend of missing values in water quality data to follow the MAR mechanism (Güler et al., 2002), the present study explores advanced imputation methods capable of handling MAR, MNAR and MCAR assumptions. By addressing the full spectrum of missing data scenarios, the study proposed a

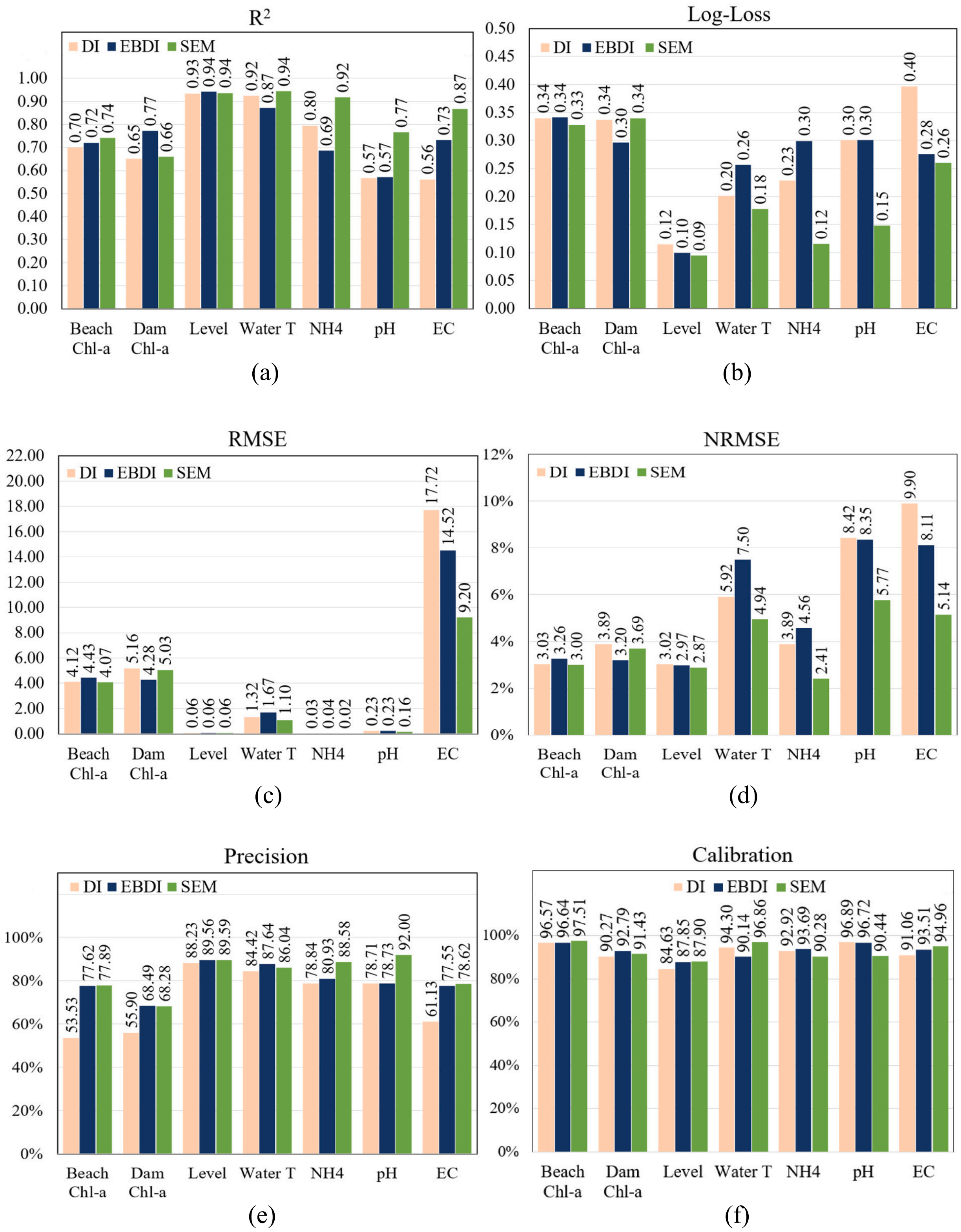


Fig. 3. Bar charts of the performance metrics (a) R^2, (b) Log-Loss, (c) RMSE, (d) NRMSE (%), (e) Precision (%), and (f) Calibration for each missing value imputation model DI, EBDI and SEM on the selected critical variables.

methodological approach for understanding water quality dynamics under fluctuating missing data scenarios. This advancement not only aligns with the latest techniques for inferring missing values but also sets a new benchmark in how water resources can be predicted, monitored, and managed amid the growing challenges posed by climate change scenarios. In the following subsections, three methods for inferring missing values are presented: Dynamic Imputation (DI), Entropy-Based Dynamic Imputation (EBDI), and Structural Expectation-Maximization (SEM), using BayesiaLab v.10.2 software.

2.4.1. *Dynamic imputation*

This method is relatively straightforward compared to EBDI and

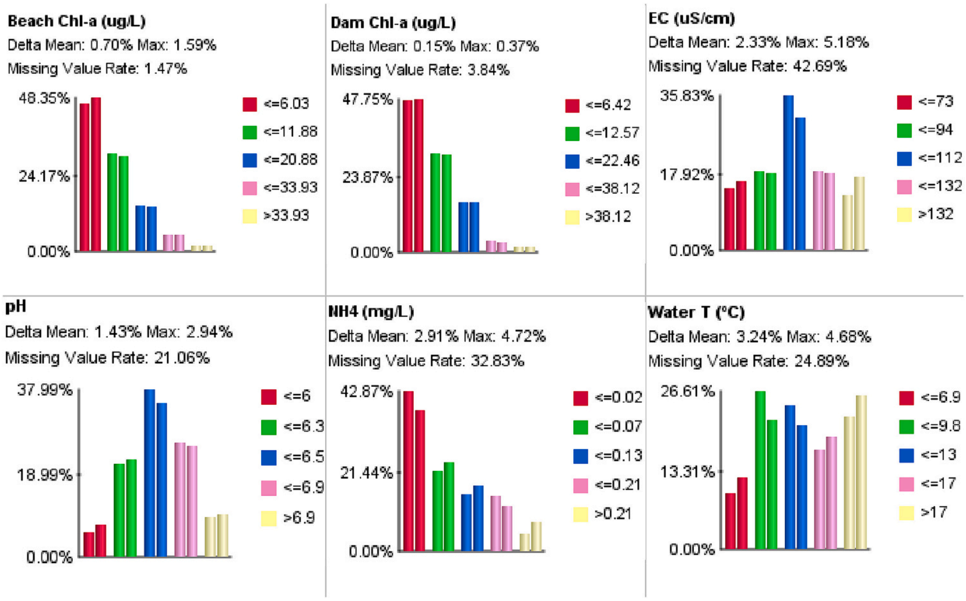


Fig. 4. Missing value imputation performance. For each color pair, the first column (on the left) corresponds to the empirical distribution of the variables, while the second column (on the right) corresponds to the probabilities calculated with the Bayesian network. The top part also shows the average of the difference between the empirical and the inferred probability (Delta Mean), the largest difference between both (Max) and the frequency of missing values for each variable in the dataset (Missing Value Rate).

SEM. It focuses on the adjustment to time-based variations in data. Although advance, DI involves fewer complex calculations than those integrating advanced statistical models. As with static imputation, initially inferred values are obtained from random extractions of the marginal distributions of observed data, which act as placeholders. However, in a second step, DI updates the missing values during the structural learning of the Bayesian network (Gámez et al., 2011). In this way, the updated network functions as a support to infer the used distribution during imputation, considering the observed variables and their previously inferred values (Conrady and Jouffle, 2015). This sets DI apart from the static method and leads to its enhanced performance. From a mathematical perspective, based on Eq. (1), the marginal likelihood can be decomposed into a single element for each local distribution:

$$P(D|G) = \int P(D|G,\theta)P(\theta|G)d\theta = \prod_{i=1}^{N} P(X_i|\Pi_{X_i},\theta_{X_i})P(\theta_{X_i}|\Pi_{X_i})d\theta_{X_i} \tag{3}$$

2.4.2. Entropy-based dynamic imputation

Entropy-based dynamic imputation is similar to dynamic imputation, but it prioritizes the imputation of missing values based on the entropy of the conditional probability distributions in the Bayesian network. The rational is to use the uncertainty in the network to guide the imputation process. In this method, the missing value with the highest entropy is inferred first, with subsequent imputations prioritized based on the decreasing entropy of the conditional probability distributions in the network (Scutari, 2018). The equation for entropy (H) computation can be expressed as follows:

$$H(X_i) = -\sum_{x \epsilon X} p(x_i) log_2(p(x_i)) \tag{4}$$

where x_i represent the potential outcomes of X_i, and their corresponding probabilities $(p(x_i))$ indicate the complexity of the network and determine the number of bits needed for its representation.

2.4.3. Structural expectation-maximization

Structural expectation-maximization involves an iterative algorithm to find the most probable explanation estimates of parameters in the Bayesian model. The basic idea is to iteratively compute the expected values of the data parameters (E-step) and then, use those inferred values to find the structure that maximizes the expected log-likelihood function (M-step) (Friedman, 1998). The specific formulas underlying the E-step $(E(D^M, D^O|G,\theta))$, and M-step $(P(\theta|G, D^O, D^M))$ of the Maximum Likelihood Estimation (MLE) algorithm are described as follows:

$$E-step : Q\left(\theta \mid \theta^{(t)}\right) = E_{X_m \sim p\left(\bullet|X_o,\theta^{(t)}\right)}[logp(X_o, X_m|\theta)] \tag{5}$$

$$M-step : \theta^{(t+1)} = arg_\theta max Q\left(\theta \mid \theta^{(t)}\right) \tag{6}$$

where $Q(\theta \mid \theta^{(t)})$ represents the expected value of the log-likelihood function of a missing parameters vector θ, regarding the current conditional distribution of a set of missing values (X_m) given a set of observed data (X_o), and the current estimates of $\theta^{(t)}$.

Finally, the general formula for structural expectation-maximization can be expressed as follows:

$$\theta^{(t+1)} = arg_\theta max E_{X_m \sim p\left(\bullet|X_o,\theta^{(t)}\right)}[logp(X_o, X_m \mid \theta)] \tag{7}$$

2.5. Network performance and validation methods

The selection of the Bayesian structural learning algorithms applied in this study has been carried out following the principle of minimum description length (MDL). The MDL allows to identify those algorithms that provide the most accurate and reliable explanation of the data. Score-based approaches, such as MDL, evaluate a function (metric) that measures the quality of a candidate network concerning the available data. The term MDL, derived from information theory and commonly used in artificial intelligence, can be defined mathematically as follows (Friedman, 1997; Conrady and Jouffle, 2015):

$$MDL(B,D) = \alpha DL(B) + DL(D|B) \tag{8}$$

where $DL(B)$ represents the number of bits needed to represent the model (graph and probabilities), $DL(D|B)$ represents the number of bits needed to represent the model data (i.e., probability of the data given

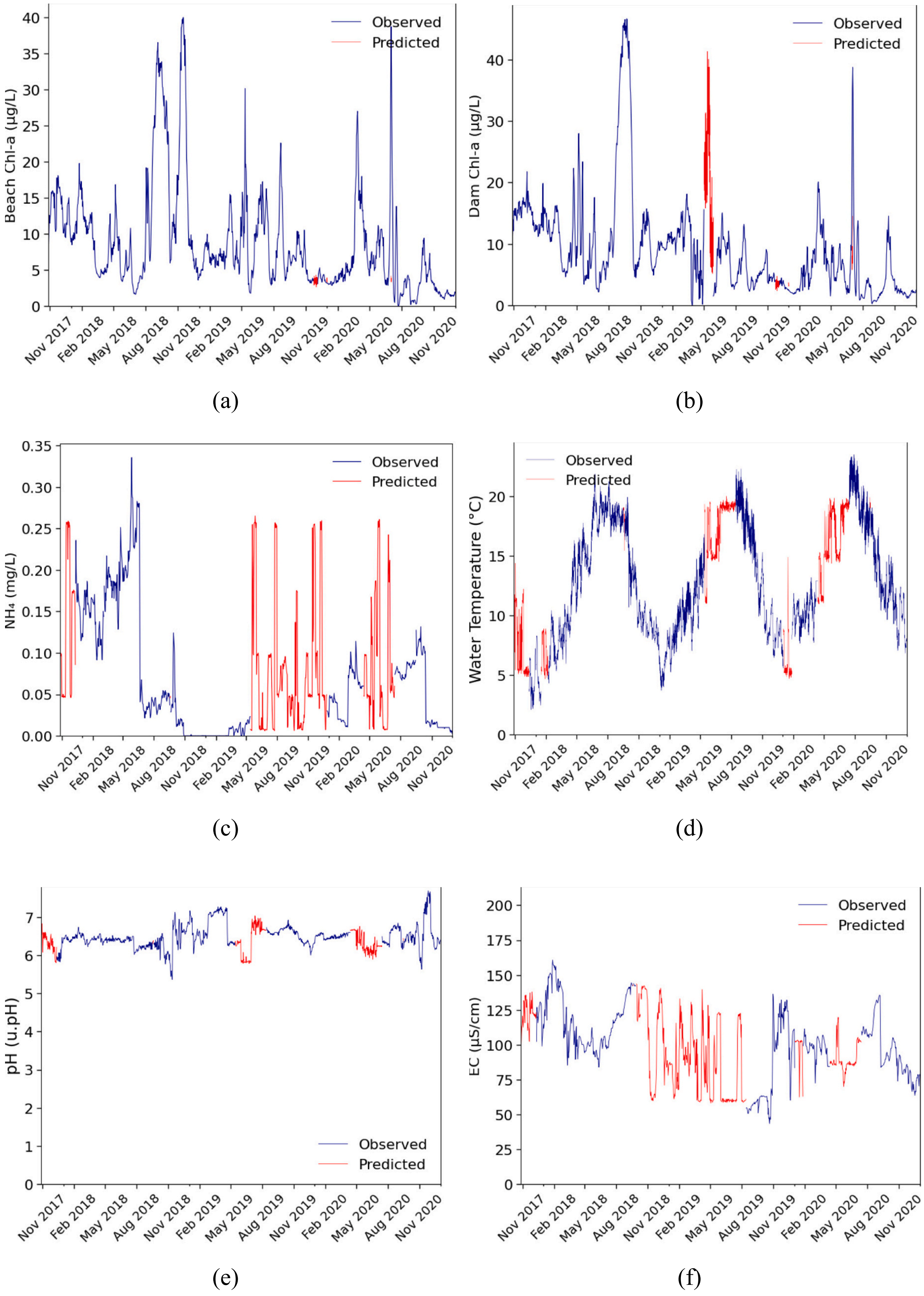


Fig. 5. Temporal representation of observed (blue) and predicted (red) values for chlorophyll concentration at (a) dam (μg/L) and (b) beach (μg/L), (c) NH_4 (mg/L), (d) water temperature (°C), (e) pH (u.pH), and (f) EC ((μS/cm). Figure created using Python and the pandas, numpy, and matplotlib libraries.

Table 4
SEM performance metrics MAE. RMSE, NRMSE (%), and R^2 of the dataset for 10 %, 15 %, and 30 % of missing values.

	MAE			RMSE			NRMSE			R^2		
	10 %	15 %	30 %	10 %	15 %	30 %	10 %	15 %	30 %	10 %	15 %	30 %
Dam Chl-a	0.46	0.85	1.35	2.88	4.43	5.05	2.15	2.83	3.78	0.95	0.89	0.86
Beach Chl-a	0.46	0.74	1.36	3.19	3.84	5.49	2.35	2.83	4.04	0.94	0.91	0.83
NH_4	0.01	0.01	0.01	0.04	0.05	0.06	4.25	5.20	7.04	0.92	0.88	0.78
PO_4^{3-}	0.02	0.03	0.04	0.30	0.36	0.46	2.81	3.31	4.22	0.88	0.84	0.74
EC	1.73	2.53	4.75	8.33	9.95	13.67	4.66	5.56	7.64	0.94	0.91	0.84
DO	0.15	0.21	0.39	0.63	0.75	1.01	5.16	6.09	8.24	0.96	0.95	0.90
pH	0.02	0.04	0.06	0.13	0.15	0.20	4.72	5.41	7.43	0.90	0.87	0.75
Water Tª	0.21	0.31	0.58	0.88	0.15	1.49	3.96	4.85	6.66	0.97	0.96	0.93
Turbidity	0.59	0.83	1.47	4.64	5.45	7.37	2.32	2.73	3.69	0.89	0.85	0.73
Water level	0.01	0.01	0.03	0.04	0.05	0.08	2.05	2.53	4.00	0.97	0.96	0.90
Water flow	0.33	0.48	1.06	2.01	2.43	4.67	1.67	2.02	3.87	0.97	0.95	0.81
Precipitation	0.03	0.04	0.07	0.43	0.47	0.63	3.15	3.44	4.65	0.88	0.85	0.73
Air temperature	0.49	0.72	1.39	0.47	2.53	3.59	4.39	5.30	7.52	0.94	0.92	0.84

the Bayesian network), and α is the structural coefficient (ranging from 0 to 150). The structural coefficient modifies the relative weighting of $DL(B)$ and $DL(D|B)$, allowing the number of observations to be adjusted.

To provide a comprehensive evaluation of model accuracy and fit, essential for validating and comparing predictive models each of them was assessed by means of the root mean square error (RMSE) (Jain and Singh, 2003), the normalize root mean squared error (NRMSE), the coefficient of determination (R^2) (Wright, 1921), and mean absolute error (MAE) (Chicco et al., 2021; Zakwan et al., 2022):

$$\mathrm{RMSE} = \sqrt{\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{X}_{\mathrm{o,i}} - \mathrm{X}_{\mathrm{p,i}}{}^{2}}, \tag{9}$$

$$0\ (\text{best value}) \le \mathrm{RMSE} < +\infty(\text{worst value})$$

$$\mathrm{NRMSE}\ (\%) = \frac{\sqrt{\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{X}_{\mathrm{o,i}} - \mathrm{X}_{\mathrm{p,i}}{}^{2}}}{\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{X}_{\mathrm{obs,i}}} \times 100, \tag{10}$$

$$0\ (\text{best value}) \le \mathrm{NRMSE} < 100\ (\text{worst value})$$

$$\mathrm{R}^2 = \frac{\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{X}_{\mathrm{o,i}} - \overline{\mathrm{X}}_{\mathrm{o}}\mathrm{X}_{\mathrm{p,i}} - \overline{\mathrm{X}}_{\mathrm{p}}{}^{2}}{\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{X}_{\mathrm{o,i}} - \overline{\mathrm{X}}_{\mathrm{o}}{}^{2} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathrm{X}_{\mathrm{p,i}} - \overline{\mathrm{X}}_{\mathrm{p}}{}^{2}}, \tag{11}$$

$$0\ (\text{worst value}) \le \mathrm{R}^2 < 1\ (\text{best value})$$

$$\mathrm{MAE} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} \left|\mathrm{X}_{\mathrm{o,i}} - \mathrm{X}_{\mathrm{p,i}}{}^{2}\right|, \tag{12}$$

$$0\ (\text{best value}) \le \mathrm{MAE} < +\infty(\text{worst value})$$

where X_o and X_p correspond to the observed and predicted values for the N observations. Likewise, $\overline{X}_o$ and $\overline{X}_p$ represent the mean of the measured and predicted variables, respectively.

On the other hand, variable discretization was also implemented into the analysis. Discretization simplifies data, aiding in model interpretability and assessment of classification accuracy and reliability. The resulting values of correctly and incorrectly classified observations were assessed based on precision, calibration, and logarithmic loss (log-loss) of the models. These evaluation parameters were defined as:

$$\text{Precision} = \frac{\text{Observations correctly classified as A}}{\text{Total observations as A}} \tag{13}$$

$$0\ (\text{worst value}) \le \text{Precision} < 100\ (\text{best value})$$

$$\text{Calibration Index} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}\left(1 - \frac{2\mathrm{R}_i}{Y_i}\right), \tag{14}$$

$$0\ (\text{worst value}) \le \text{Calibration} < 100\ (\text{best value})$$

where R_i represents the mis-calibration area for each calibration curve, Y_i are the respective areas of non-gray (total possible area) below the diagonal for each calibration curve.

Finally, the logarithmic loss returned by the model was calculated based on:

$$\text{Log} - \text{Loss} = \frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}} log_2(P_{T_i}) \tag{15}$$

$$0\ (\text{best value}) \le \text{Log} - \text{Loss} < 1\ (\text{worst value})$$

where P_{T_i} represents the posterior probability provided by the model for the true state of the Target Node in the i^{th} row of the dataset. The logarithmic loss (Log-Loss) imposes significant penalties on inaccurate predictions made with a high level of certainty.

3. Results and discussion

3.1. Unsupervised structural learning

As first step in the dynamic Bayesian network-based processing system, the entire portfolio of unsupervised learning algorithms available in the BayesiaLab v.10.2 software was tested. The aim was to select the algorithm that could describe the data most efficiently, using the MDL metric. To select the best algorithm, MDL values for each algorithm were compared, and the lowest value was considered. As a result, the SopLEQ was selected for DI (3,145,272.40) and SEM (3,115,917.14), and the Taboo (3,153,398.79) was chosen for EBDI. In Table 2 it is shown a comparison of the different algorithms tested and their computational performance. This approach ensures that the most efficient algorithms are chosen for each dynamic imputation processes.

In Fig. 2, a radial layout is presented, showcasing an unsupervised Bayesian network developed through the SopLEQ algorithm. This network integrates the SEM method for the inference of missing values. From an exploratory standpoint, this radial layout offers a well-organized visual representation, highlighting the complex interconnections within the network. Additionally, a thorough exploratory analysis of this model reveals the pivotal role of the temporal variable month as a central node, underlining its significant influence in the Limia river features.

3.2. Assessment of the dynamic Bayesian models

In this section, the performance and predictive capacity of the

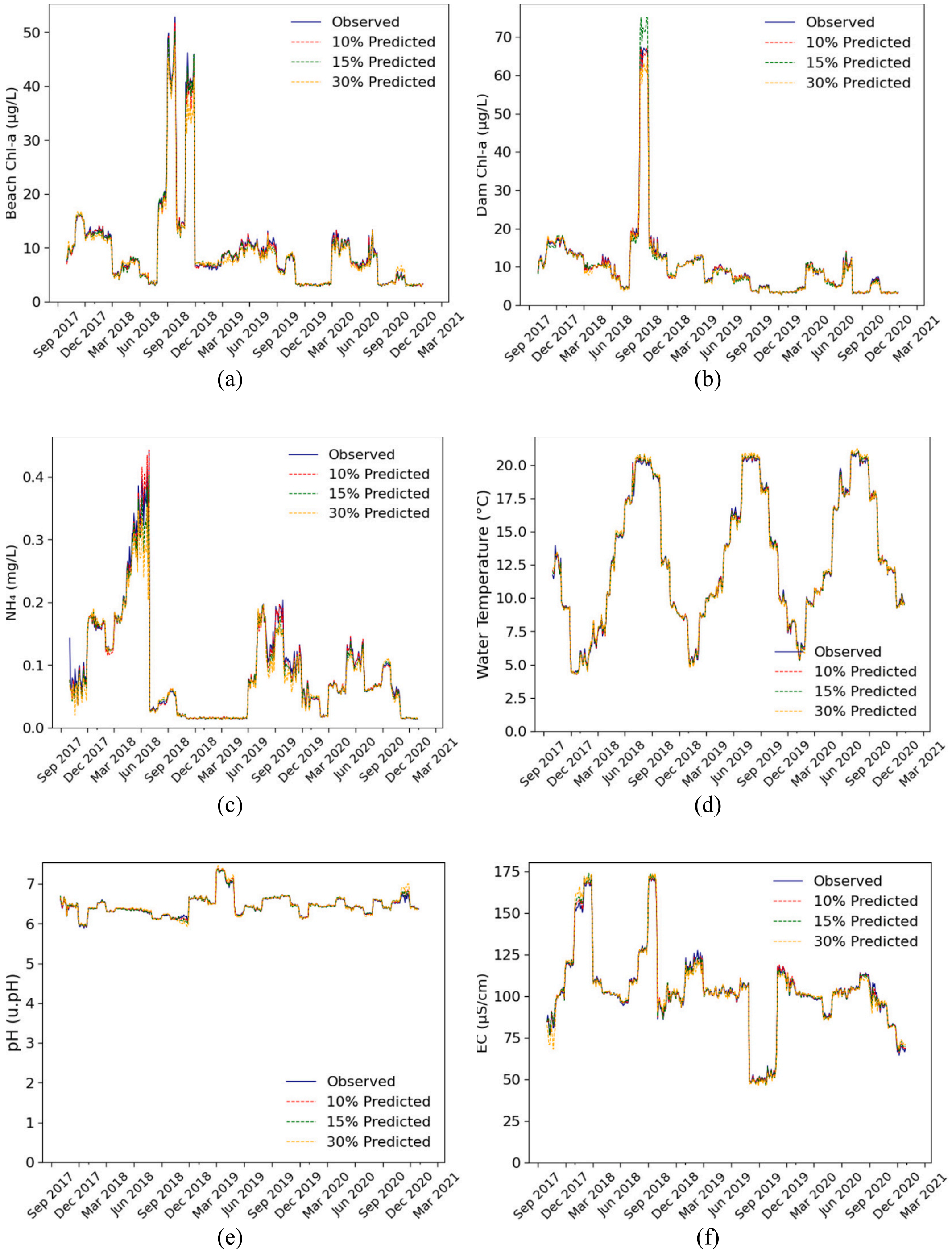
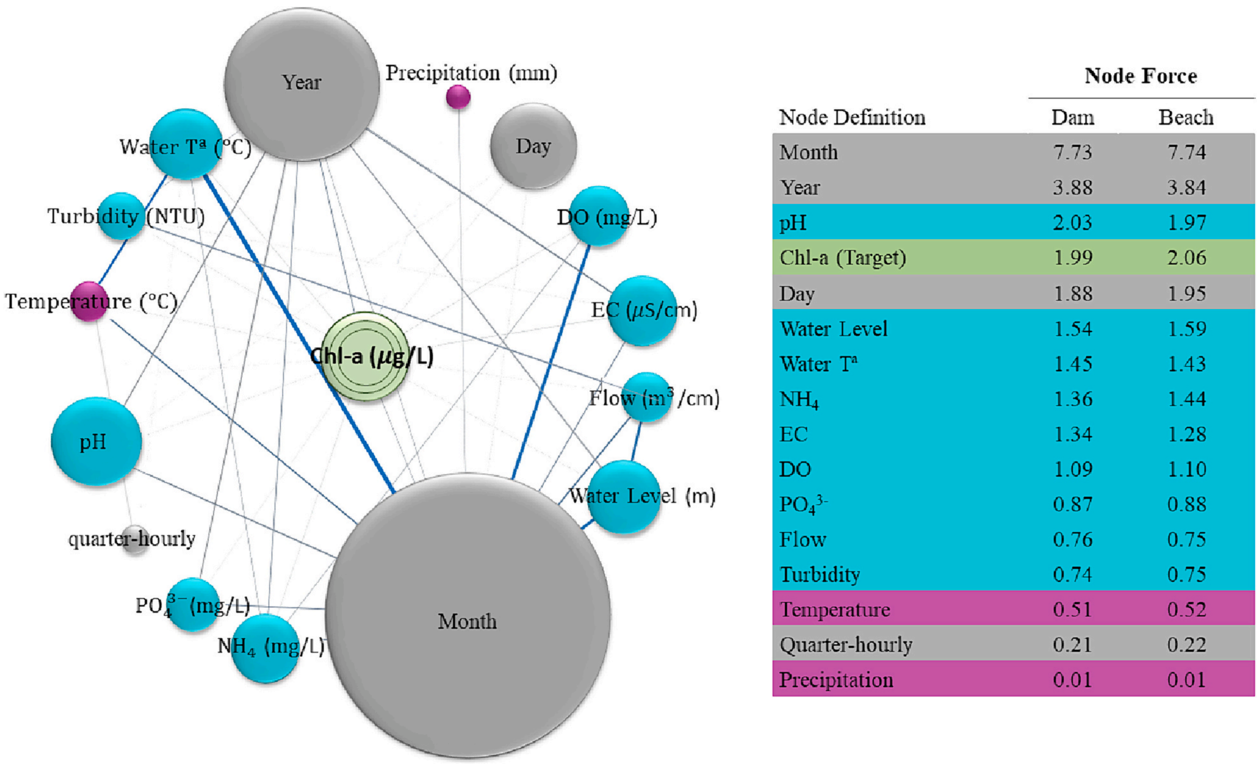


Fig. 6. Representation of observed (dotted blue) and predicted values for 10 % of MV (dotted red), 15 % of MV (dotted green), 30 % of MV (dotted orange) in the database. Analyzed variables: chlorophyll concentration at (a) dam (μg/L) and (b) beach (μg/L), (c) NH_4 (mg/L), (d) water temperature (°C), (e) pH (u. pH), and (f) EC (μS/cm). Figure created using Python and the pandas, numpy, and matplotlib libraries.

missing value imputation processes DI, EBDI, and SEM are evaluated. Table 3 presents the metrics utilized to define the models' goodness of fit, considering the imputation of missing values in the eleven physicochemical water and hydraulic parameters measured by the SAICA and SAIH sensors, as well as the Chl-a concentration recorded by the buoys located at the beach and reservoir dam.

The results show that the dynamic processing of missing values, SEM, outperformed the DI and EBDI methods based on the evaluated metrics. It is important to note that R^2, RMSE, and NRMSE (%) are commonly used metrics for evaluating model fit and offer unique information that allows analyzing different aspects of the model. In this case, the results obtained using the SEM methodology indicate a higher R^2 value (0.76) which makes it potentially the best model to explain the variability of the data. In addition, the lowest RMSE value (2.92) was observed, indicating that the inferred missing values are closer to the observed ones. Similarly, the SEM model provided the lowest NRMSE (3.85 %), implying a better accuracy in the predicted values relative to the range of the dependent variable. Otherwise, according to the precision (80 %), log-loss (0.22), and calibration (92.79 %) results of the SEM model, it is possible to observe that SEM algorithm corresponds to the best performance. Even so, despite EBDI and DI showed lower precision and log-loss, both provided acceptable results (Table 3).



Node Definition	Node Force Dam	Node Force Beach
Month	7.73	7.74
Year	3.88	3.84
pH	2.03	1.97
Chl-a (Target)	1.99	2.06
Day	1.88	1.95
Water Level	1.54	1.59
Water Tª	1.45	1.43
NH$_4$	1.36	1.44
EC	1.34	1.28
DO	1.09	1.10
PO$_4^{3-}$	0.87	0.88
Flow	0.76	0.75
Turbidity	0.74	0.75
Temperature	0.51	0.52
Quarter-hourly	0.21	0.22
Precipitation	0.01	0.01

Fig. 7. Supervised Bayesian network created with the Augmented Naive Bayes algorithm. The colors of the network nodes correspond to the classification of the type of variables measured, while their sizes represent the force of the nodes (values also represented in the adjacent table). Also, the arcs highlighted in blue connect the variables with more mutual information.

Table 5
Results of the marginal probabilities and the Relative Binary Mutual Information (RBMI) analysis with Chl-a target states.

	No risk		Precautionary alarm		WHO medium alarm		WHO high alarm	
	Dam	Beach	Dam	Beach	Dam	Beach	Dam	Beach
Marginal probability	*37.29 %*	*38.33 %*	*28.78 %*	*32.14 %*	*32.50 %*	*28.99 %*	*1.43 %*	*0.25 %*
[a]Chl-a (µg/L)	32.01 %		5.59 %		28.11 %		16.65 %	
Date (dd/mm/yy)	26.37 %	18.20 %	11.19 %	13.04 %	20.79 %	14.04 %	43.38 %	25.27 %
Water Level (m)	1.14 %	1.57 %	1.98 %	3.73 %	0.37 %	4.49 %	5.79 %	3.30 %
EC (µS/cm)	8.18 %	10.00 %	4.23 %	3.12 %	12.78 %	14.93 %	25.01 %	7.33 %
pH (u.pH)	0.92 %	1.8 %	1.01 %	2.97 %	2.88 %	4.56 %	9.28 %	1.99 %
NH$_4$ (mg/L)	7.36 %	2.37 %	2.18 %	3.12 %	4.17 %	3.58 %	4.84 %	3.64 %
Water temperature (°C)	3.20 %	1.86 %	1.15 %	1.81 %	4.49 %	2.60 %	10.32 %	2.25 %

[a] RBMI value concerning beach buoy Chl-a concentration and vice versa.

3.3. Feature selection and prediction performance

This section focuses on the performance results of the DI, EBDI and SEM models to the dynamic imputation of missing values in each measured variable. For this purpose, a general unsupervised Bayesian analysis was performed to identify those physicochemical or hydraulic parameters with significant impact on the river system. The node force values of the SAICA and SAIH parameters were calculated for the three dynamic imputation processes to identify those variables with the greatest relative importance in the behavioral patterns of the aquatic ecosystem. The results of the node weight considering the mean values of the three models identified river level (2.53), water temperature (2.42), and NH$_4$ water concentration (2.23) as the three most significant variables, followed by pH (1.73) and EC (1.62).

Therefore, based on the analysis of the underlying data structure and the issue of cyanobacterial proliferation in the As Conchas reservoir, Fig. 3 presents the performance of the three data imputation methods (DI, EBDI, and SEM) on the significant identified variables. Fig. 3 was created using Python, with the employment of pandas, numpy, and matplotlib libraries for data handling and visualization generation. The evaluation metrics applied in the analysis are presented relatively to R^2 (Fig. 3a), logarithmic loss (Fig. 3b), RMSE (Fig. 3c), NRMSE (Fig. 3d), precision (Fig. 3e) and calibration (Fig. 3f).

Across all measured variables, the three missing value imputation methods show comparable R^2 values. However, for certain variables such as NH$_4$, EBDI exhibits a significantly lower R^2 value (0.69) compared to DI (0.80) and SEM (0.92). Similarly, in the case of EC, DI (0.56) also shows a significantly lower value than EBDI (0.73) and SEM (0.87). These critical variables present a more consistent imputation of missing values with SEM model, as it achieves higher R^2 values for Beach Chl-a (0.74), EC (0.87), NH$_4$ (0.92), water level (0.94), water Temperature (0.94), and pH (0.77). This indicates a better model fit to the data and a higher ability to explain the variability of the measured parameters. Relatively to Log-Loss results, it is noteworthy that lower values of Log-Loss indicate better performance. Algorithmic loss penalizes inaccurate predictions, especially with high certainty conditions. Regarding Log-Loss values, as with the previously mentioned statistical metrics, SEM potentially performs better. Otherwise, RMSE and NRMSE (%)

values suggest that data imputation models exhibit a good fit, with values below the commonly NRMSE accepted threshold of 10 %.

Finally, the SEM imputation algorithm shows outstanding performance in several aspects, followed by the EBDI. Notably, a significant improvement in the imputation of the EC variable is observed with the SEM model, reducing RMSE, NRMSE and Log-Loss values by >50 % compared to the DI model. This imputation algorithm also demonstrated robust calibration on almost variables, indicating that the predicted probabilities are consistent with the observed frequencies. In addition, SEM also obtained remarkably high accuracy in pH (92), NH_4 (88.58), water level (89.59), and water temperature (86.04) indicating that the inferred values were correctly classified. However, the accuracy values are reduced in the Dam Buoy zone, probably due to the greater water depth than in the Beach Buoy zone (>10 m). This may result in wider oscillations of Chl-a occurrence, and consequently a more difficult and accurate prediction.

As a further step in the analysis of missing value imputation, six adjacent histograms are presented for each variable using the SEM missing value imputation model (Fig. 4). These histograms depict the distribution of observations (left column) and the distribution inferred by the learned Bayesian network (right column). By analyzing the variable marginal distribution, it can be observed that the water flow rate presented no bias introduced by the missingness (0.01 % missing values). However, variables such as water temperature, pH, EC, Beach Chl-a and to a lesser extent Dam Chl-a were overestimated, i.e. the lower values were missed, and the current probability distribution was recovered. In the case of water NH_4 concentration, the empirical distribution of the lowest values was being underestimated. The visualization of the variables marginal probability distribution is a powerful analytical tool, as it allows to understand the extent of missing values have biased the data, as well as to identify the presence of MNAR or MAR mechanisms in these biases. This information is crucial for researchers on make decision processes based on accurate and representative data.

Regarding the inferred missing values, the values of the last updated network have been extracted (Fig. 5).

3.4. *Dataset validation*

In the last step of the analysis, new test data are generated to validate the performance of the SEM algorithm. This involves the creation of a new dataset that aligns with the Joint Probability Distribution (JPD) encoded by the reference network. In this way, it ensures that the test data accurately represent the underlying distribution captured by the model. This approach is crucial to maintaining the validity and relevance of the evaluation process. Finally, 50,000 data points were generated and tested for 5 %, 15 %, and 30 % missing values (Table 4).

The values of the most recent updated network have been extracted to showcase the inferred missing values (Fig. 6).

3.5. *Supervised Bayesian analysis*

Finally, this section demonstrates the applicability of the methodology proposed in this study by identifying the variables significantly associated with the risk of exceeding the WHO chlorophyll-a concentration threshold (Chorus and Welker, 2021) in the most problematic areas for the occurrence of cyanobacterial blooms.

For this purpose, the values of chlorophyll-a concentration have been discretised into four alarm levels denoted as: no risk (<5 μg/L), precautionary alarm (5 to 10 μg/L), WHO medium alarm (10 to 50 μg/L), and WHO high alarm (>50 μg/L). Considering that microalgal blooms do not necessarily occur in the same regions of the reservoir, nor with the same intensity, two supervised Bayesian networks were constructed to investigate the statistical association between model variables and chlorophyll-a concentration states (target node) as a function of record location (Fig. 7).

The analysis of node force between the target nodes and other variables in the model demonstrated a similar behavior at the dam and the beach (Fig. 7). The results revealed that temporal variables exerted the greatest influence on the model, with the variable month as the most significant predictor (7.73–7.74). Following this, the next most highly ranked predictor variable was pH (1.97–2.03). As highlighted in the study conducted by Acuña-Alonso et al. (2020), the optimal range for cyanobacteria growth occurs to water pH value between 6 and 9 (neutral or alkaline environment). Notably, 94.43 % of recorded pH water values were within this interval, indicating a conducive environment for the rapid cyanobacteria growth. In fact, pH has also been identified as the most influential variable on cyanobacteria behavior in the reservoir, as demonstrated by Mozo et al. (2022).

The second most significant water quality variable corresponds to water level (1.54–1.59), followed by water temperature (1.43–1.45), NH_4 (1.36–1.44) and EC (1.28–1.34). Water level variation is a critical factor directly related to water flow, and some studies indicate that cyanobacteria dominance is commonly found in reservoirs and slow-flowing rivers (Xu et al., 2023). In addition, water temperature is one of the most extensively studied variable in microalgae growth (Sarma, 2013; Huisman et al., 2018). Water temperature has a significant influence on the growth of cyanobacteria, with an optimal temperature between 15 °C and approximately 28 °C (Robarts and Zohary, 1987; Acuña-Alonso et al., 2020).

In summary, the most influence parameters on the proliferation of cyanobacteria point to be water electrical conductivity and ammonium concentration as crucial parameters to control and determine water quality (DOUE-L-2000-82524). Previous studies have analyzed the influence of the two water parameters on cyanobacterial proliferation in the As Conchas Reservoir (di Blasi et al., 2013; Garzon-Vidueira et al., 2020), revealing significant diffuse pollution issues in the area and a strong relationship with human activity. In addition, these parameters can be influenced by substances dragged from adjacent lands, such as ammonium occurrence associated to fertilizers or salts that will increase water electrical conductivity.

To a better understanding of water Chl-a concentration and distribution, a second level of analysis was carried out to assess the relevance of the six predictors that will contribute to an higher risk of exceeding the Chl-a threshold alarm levels: no risk (<5 μg/L), precautionary alarm (5 to 10 μg/L), WHO medium alarm (10 to 50 μg/L), and WHO high alarm (>50 μg/L; Table 5). The local impact analyses enabled the quantification of the amount of information associated to each variable on the Chl-a concentration range.

Previous studies have shown that the presence and proliferation of cyanobacteria in surface water are clearly influenced by seasonal factors, such as air temperature, sunlight, and nutrient availability (Iglesias et al., 2016). These factors are fundamental on the dynamics of Chl-a and their relationship with water quality parameters. On the study area is possible to identify the pattern of dependence and simultaneously quantify the mutual information shared among these factors (Table 5). In particular, measurement date, electrical conductivity, water temperature and ammonium concentration show a high percentage of mutual information for high Chl-a concentration (high alarm). The presence of nutrients in the river, which influences other factors such as electrical conductivity should be noted as the main source of nitrates in the As Conchas River appears to be organic inputs to the soil through livestock waste (Garzon-Vidueira et al., 2020). This information supports the major contribution to water electrical conductivity values and ammonium concentration as a relevant risk to cyanobacterial contamination in the reservoir, especially at concentration higher than 10 μg/L.

4. Conclusions

The impact of missing data on the reliability of water quality analyses can be managed. This study introduced a methodological approach that addresses and integrates missing information to advance scientific understanding of water quality. This issue is particularly relevant, as in the present research, where missing data of varying magnitude was impeding to ascertain the root causes associated to water cyanobacteria proliferation in the case study of Limia river in Spain. The present research demonstrated that a methodological approach based on Bayesian dynamic imputation allows to reliably approximate observed values in the missing water samples from the monitoring sensors. These methods showed precise estimations and adequate calibration, indicating their effectiveness in improving the behavior of these aquatic ecosystems. The analyses conducted, considering the entire database or individual variables – Chl-a concentration (μg/L), river level (m), water temperature (°C), NH_4 (mg/L), pH (u. pH), and EC (μS/cm) – showed that SEM imputation method outperformed the DI and EBDI methods also applied, although EBDI and DI demonstrated good performance. Looking further, it is still necessary to understand to what extent this proposed system can handle large volumes of missing data in a reliable manner.

Finally, the applicability of this methodology was shown by showcasing cyanobacteria concentration in the As Conchas reservoir, where it had almost reached alert level 2 according to World Health Organization (WHO) guidelines (Moron-Lopez et al., 2021; Chorus and Welker, 2021), and resulting in water supply cuts for the population. The Chl-a water concentration was categorized into four alarm levels: no risk (<5 μg/L), precautionary alarm (5 to 10 μg/L), WHO medium alarm (10 to 50 μg/L), and WHO high alarm (>50 μg/L). A supervised Bayesian methodology implemented shed light on the problem, revealing that temporal variables exerted the greatest influence on cyanobacterial proliferation, with RMBI values up to 43 % concerning the identification of a maximum cyanobacterial water contamination alarm situation. These findings and practical demonstration contribute to the scientific assessment of water quality subject to varying missing data.

CRediT authorship contribution statement

M. Pazo: Writing – original draft, Visualization, Software, Methodology, Formal analysis, Conceptualization. **S. Gerassis:** Writing – original draft, Validation, Methodology, Formal analysis, Conceptualization. **M. Araújo:** Writing – review & editing, Validation, Supervision, Investigation, Funding acquisition, Conceptualization. **I. Margarida Antunes:** Writing – review & editing, Validation. **X. Rigueira:** Writing – review & editing, Validation, Resources, Investigation, Data curation.

Declaration of competing 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 paper.

Data availability

Data will be made available on request.

Acknowledgements

This research was funded by the Spanish Ministry of Science and Innovation (MCIN/AEI/10.13039/501100011033), grant number PID2020-116013RB-I00. This work is developed within the activities of the FCT - Foundation for Science and Technology, I.P., projects UIDB/04683/2020 and UIDP/04683/2020.

References

Abdelkarim, B., Antunes, I.M.H.R., Abaab, N., Agoubi, B., 2023. Modeling groundwater recharge mechanisms in semi-arid regions: integration of hydrochemical and isotopic data. Euro-Mediterr. J. Environ. Integr. https://doi.org/10.1007/s41207-023-00400-3.

Acock, A.C., 2005. Working with missing values. J. Marriage Fam. 67 (4), 1012–1028. https://doi.org/10.1111/J.1741-3737.2005.00191.X.

Acuña-Alonso, C., Lorenzo, O., Álvarez, X., Cancela, Á., Valero, E., Sánchez, Á., 2020. Influence of Microcystis sp. and freshwater algae on pH: changes in their growth associated with sediment. Environ. Pollut. 263, 114435 https://doi.org/10.1016/J.ENVPOL.2020.114435.

Acuña-Alonso, C., Álvarez, X., Lorenzo, O., Cancela, Á., Valero, E., Sánchez, Á., 2021. Water toxicity in reservoirs after freshwater algae harvest. J. Clean. Prod. 283, 124560 https://doi.org/10.1016/J.JCLEPRO.2020.124560.

Albers, C., 2023. Water Quality Data Pathfinder - Find Data | Earthdata. October 12. https://www.earthdata.nasa.gov/learn/pathfinders/water-quality-data-pathfinder/find-data#runoff.

Al-Sabhan, W., Mulligan, M., Blackburn, G.A., 2003. A real-time hydrological model for flood prediction using GIS and the WWW. Comput. Environ. Urban. Syst. 27 (1), 9–32. https://doi.org/10.1016/S0198-9715(01)00010-2.

Asif, Z., Chen, Z., Sadiq, R., Zhu, Y., 2023. Climate change impacts on water resources and sustainable water management strategies in North America. Water Resour. Manag. 37 (6–7), 2771–2786. https://doi.org/10.1007/S11269-023-03474-4/FIGURES/1.

Boelee, E., Geerling, G., van der Zaan, B., Blauw, A., Vethaak, A.D., 2019. Water and health: from environmental pressures to integrated responses. Acta Trop. 193, 217–226. https://doi.org/10.1016/J.ACTATROPICA.2019.03.011.

Carballeira, R., Vieira, D.N., Febrero-Bande, M., Muñoz Barús, J.I., 2018. A valid method to determine the site of drowning. Int. J. Leg. Med. 132 (2), 487–497. https://doi.org/10.1007/S00414-017-1708-1/TABLES/2.

Chen, J., Zhong, P.A., An, R., Zhu, F., Xu, B., 2019. Risk analysis for real-time flood control operation of a multi-reservoir system using a dynamic Bayesian network. Environ. Model. Software 111, 409–420. https://doi.org/10.1016/J.ENVSOFT.2018.10.007.

Chicco, D., Warrens, M.J., Jurman, G., 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput. Sci. 7, 1–24. https://doi.org/10.7717/PEERJ-CS.623.

Chorus, I., Welker, M. (Eds.), 2021. Toxic Cyanobacteria in Water: A Guide to Their Public Health Consequences, Monitoring and Management, 2nd ed. CRC Press. https://doi.org/10.1201/9781003081449.

Conrady, S., Jouffle, L., 2015. Bayesian Netwroks & BayesialLab – A Practical Introduction for Researchers. Franklin, TN, Bayesia USA (ISBN: 978-0-996533-0-0).

Desbureaux, S., Mortier, F., Zaveri, E., van Vliet, M.T.H., Russ, J., Rodella, A.S., Damania, R., 2022. Mapping global hotspots and trends of water quality (1992–2010): a data driven approach. Environ. Res. Lett. 17 (11) https://doi.org/10.1088/1748-9326/AC9CF6.

Dhaoui, O., Agoubi, B., Antunes, I.M.H.R., Tlig, L., Kharroubi, A., 2022. Groundwater quality for irrigation in an arid region – application of fuzzy logic techniques. Environ. Sci. Pollut. Res. 30, 29773–29789. https://doi.org/10.1007/s11356-022-24334-5.

di Blasi, J.I.P., Martínez Torres, J., García Nieto, P.J., Alonso Fernández, J.R., Díaz Muñiz, C., Taboada, J., 2013. Analysis and detection of outliers in water quality parameters from different automated monitoring stations in the Miño river basin (NW Spain). Ecol. Eng. 60, 60–66. https://doi.org/10.1016/J.ECOLENG.2013.07.054.

Dong, Y., Peng, C.Y.J., 2013. Principled missing data methods for researchers. SpringerPlus 2 (1), 1–17. https://doi.org/10.1186/2193-1801-2-222/TABLES/3.

DOUE-L-2000-82524 Directiva 2000/60/CE del Parlamento Europeo y del Consejo, de 23 de octubre de 2000. por la que se establece un marco comunitario de actuación en el ámbito de la política de aguas. Retrieved June 06, 2023, from. https://www.boe.es/buscar/doc.php?id=DOUE-L-2000-82524.

Economic and Social Council, 2022. Meeting of the Parties to the Protocol on Water and Health to the Convention on the Protection and Use of Transboundary Watercourses and International Lakes (6th meeting; Palais des Nations, Geneva, 16–18 November).

European Commission, Directorate-General for Research and Innovation, 2022. Strategic Research and Innovation Agenda (SRIA) of the European Open Science Cloud (EOSC). Publications Office. https://doi.org/10.2777/935288.

Evans, A.E., Mateo-Sagasta, J., Qadir, M., Boelee, E., Ippolito, A., 2019. Agricultural water pollution: key knowledge gaps and research needs. Curr. Opin. Environ. Sustain. 36, 20–27. https://doi.org/10.1016/J.COSUST.2018.10.003.

Ezzati, G., Kyllmar, K., Barron, J., 2023. Long-term water quality monitoring in agricultural catchments in Sweden: impact of climatic drivers on diffuse nutrient loads. Sci. Total Environ. 864 https://doi.org/10.1016/J.SCITOTENV.2022.160978.

Friedman, N., 1997. Learning belief networks in the presence of missing values and hidden variables. In: Fisher, O.H. (Ed.), Proceedings of the Fourteenth International Conference on Machine Learning (ICML '97). Morgan Kaufmann, pp. 125–133.

Friedman, N., 1998. The Bayesian structural EM algorithm. In: Cooper, G.F., Moral, S. (Eds.), Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI-98). Morgan Kaufmann, pp. 129–138.

Gámez, J.A., Mateo, J.L., Puerta, J.M., 2011. Learning Bayesian networks by hill climbing: efficient methods based on progressive restriction of the neighborhood. Data Min. Knowl. Disc. 22 (1), 106–148. https://doi.org/10.1007/s10618-010-0178-6.

Garzon-Vidueira, R., Rial-Otero, R., Garcia-Nocelo, M.L., Rivas-Gonzalez, E., Moure-Gonzalez, D., Fompedriña-Roca, D., Vadillo-Santos, I., Simal-Gandara, J., 2020. Identification of nitrates origin in Limia river basin and pollution-determinant factors. Agric. Ecosyst. Environ. 290 https://doi.org/10.1016/J.AGEE.2019.106775.

Gerassis, S., Giráldez, E., Pazo-Rodríguez, M., Saavedra, Á., Taboada, J., 2021. Ai approaches to environmental impact assessments (Eias) in the mining and metals sector using automl and bayesian modeling. Appl. Sci. (Switzerland) 11 (17). https://doi.org/10.3390/app11177914.

Güler, C., Thyne, G.D., McCray, J.E., Turner, A.K., 2002. Evaluation of graphical and multivariate statistical methods for classification of water chemistry data. Hydrgeol. J. 10 (4), 455–474. https://doi.org/10.1007/S10040-002-0196-6.

Heckerman, D., Geiger, D., Chickering, D.M., 1995. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data, vol. 20.

Ho, K.C., Teow, Y.H., Sum, J.Y., Ng, Z.J., Mohammad, A.W., 2021. Water pathways through the ages: integrated laundry wastewater treatment for pollution prevention. Sci. Total Environ. 760, 143966 https://doi.org/10.1016/J.SCITOTENV.2020.143966.

Huang, I.H., Chang, M.J., Lin, G.F., 2022. An optimal integration of multiple machine learning techniques to real-time reservoir inflow forecasting. Stoch. Env. Res. Risk A. 36 (6), 1541–1561. https://doi.org/10.1007/S00477-021-02085-Y/FIGURES/14.

Huisman, J., Codd, G.A., Paerl, H.W., Ibelings, B.W., Verspagen, J.M.H., Visser, P.M., 2018. Cyanobacterial blooms. Nat. Rev. Microbiol. 16 (8), 471–483. https://doi.org/10.1038/s41579-018-0040-1, 2018 16:8.

Iglesias, C., Sancho, J., Piñeiro, J.I., Martínez, J., Pastor, J.J., Taboada, J., 2016. Shewhart-type control charts and functional data analysis for water quality analysis based on a global indicator. Desalin. Water Treat. 57 (6), 2669–2684. https://doi.org/10.1080/19443994.2015.1029533.

Ijadi Maghsoodi, A., Torkayesh, A.E., Wood, L.C., Herrera-Viedma, E., Govindan, K., 2023. A machine learning driven multiple criteria decision analysis using LS-SVM feature elimination: sustainability performance assessment with incomplete data. Eng. Appl. Artif. Intel. 119, 105785 https://doi.org/10.1016/J.ENGAPPAI.2022.105785.

Jain, S.K., Singh, V.P., 2003. Statistical techniques for data analysis. Dev. Water Sci. 51 (C), 207–276. https://doi.org/10.1016/S0167-5648(03)80058-8.

Johnson, Z.C., Johnson, B.G., Briggs, M.A., Snyder, C.D., Hitt, N.P., Devine, W.D., 2021. Heed the data gap: guidelines for using incomplete datasets in annual stream temperature analyses. Ecol. Indic. 122, 107229 https://doi.org/10.1016/J.ECOLIND.2020.107229.

Karimidastenaei, Z., Avellán, T., Sadegh, M., Kløve, B., Haghighi, A.T., 2022. Unconventional water resources: global opportunities and challenges. Sci. Total Environ. 827, 154429 https://doi.org/10.1016/J.SCITOTENV.2022.154429.

Lin, T.H., 2010. A comparison of multiple imputation with EM algorithm and MCMC method for quality of life missing data. Qual. Quant. 44 (2), 277–287. https://doi.org/10.1007/S11135-008-9196-5/METRICS.

Lin, W.C., Tsai, C.F., 2020. Missing value imputation: a review and analysis of the literature (2006–2017). Artif. Intell. Rev. 53 (2), 1487–1509. https://doi.org/10.1007/S10462-019-09709-4/FIGURES/8.

Little, R.J., Rubin, D.B., 2002. Single imputation methods. In: Statistical Analysis With Missing Data, pp. 59–74. https://doi.org/10.1002/9781119013563.ch4.

Little, R.J.A., Rubin, D.B., 2019. Statistical analysis with missing data. In: Statistical Analysis with Missing Data, pp. 1–449. https://doi.org/10.1002/9781119482260.

Mishra, R.K., 2023. Fresh water availability and its global challenge. Brit. J. Multidiscip. Adv. Stud. 4 (3), 1–78. https://doi.org/10.37745/BJMAS.2022.0208.

Moron-Lopez, J., Rodriguez-Sanchez, M.C., Carreno, F., Vaquero, J., Pompa-Pernia, A.G., Mateos-Fernandez, M., Pascual Aguilar, J.A., 2021. Implementation of smart buoys and satellite-based systems for the remote monitoring of harmful algae bloom in inland waters. IEEE Sensors J. 21 (5), 6990–6997. https://doi.org/10.1109/JSEN.2020.3040139.

Mozo, A., Morón-López, J., Vakaruk, S., Pompa-Pernía, Á.G., González-Prieto, Á., Aguilar, J.A.P., Gómez-Canaval, S., Ortiz, J.M., 2022. Chlorophyll soft-sensor based on machine learning models for algal bloom predictions. Sci. Reports 12 (1), 1–23. https://doi.org/10.1038/s41598-022-17299-5, 2022 12:1.

Muzammil, M., Zahid, A., Farooq, U., Saddique, N., Breuer, L., 2023. Climate change adaptation strategies for sustainable water management in the Indus basin of Pakistan. Sci. Total Environ. 878, 163143 https://doi.org/10.1016/J.SCITOTENV.2023.163143.

Ngouna, R.H., Ratolojanahary, R., Medjaher, K., Dauriac, F., Sebilo, M., Junca-Bourié, J., 2020. A data-driven method for detecting and diagnosing causes of water quality contamination in a dataset with a high rate of missing values. Eng. Appl. Artif. Intel. 95, 103822 https://doi.org/10.1016/J.ENGAPPAI.2020.103822.

O'Neil, J.M., Davis, T.W., Burford, M.A., Gobler, C.J., 2012. The rise of harmful cyanobacteria blooms: the potential roles of eutrophication and climate change. Harmful Algae 14, 313–334. https://doi.org/10.1016/J.HAL.2011.10.027.

Özerol, G., Dolman, N., Bormann, H., Bressers, H., Lulofs, K., Böge, M., 2020. Urban water management and climate change adaptation: a self-assessment study by seven midsize cities in the North Sea Region. Sustain. Cities Soc. 55, 102066 https://doi.org/10.1016/J.SCS.2020.102066.

Paul, P.K., Zhang, Y., Ma, N., Mishra, A., Panigrahy, N., Singh, R., 2021. Selecting hydrological models for developing countries: perspective of global, continental, and country scale models over catchment scale models. J. Hydrol. 600, 126561 https://doi.org/10.1016/J.JHYDROL.2021.126561.

Pearl, J., 1988. Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, San Mateo, CA.

Pierrat, É., Laurent, A., Dorber, M., Rygaard, M., Verones, F., Hauschild, M., 2023. Advancing water footprint assessments: combining the impacts of water pollution and scarcity. Sci. Total Environ. 870 https://doi.org/10.1016/J.SCITOTENV.2023.161910.

Rigueira, X., Pazo, M., Araújo, M., Gerassis, S., Bocos, E., 2023. Bayesian machine learning and functional data analysis as a two-fold approach for the study of acid mine drainage events. Water (Switzerland) 15 (8). https://doi.org/10.3390/w15081553.

Robarts, R.D., Zohary, T., 1987. Temperature effects on photosynthetic capacity, respiration, and growth rates of bloom-forming cyanobacteria. N. Z. J. Mar. Freshw. Res. 21, 391–399. https://doi.org/10.1080/00288330.1987.9516235.

Roy, A., Murtugudde, R., Narvekar, P., Sahai, A.K., Ghosh, S., 2023. Remote sensing and climate services improve irrigation water management at farm scale in Western-Central India. Sci. Total Environ. 879, 163003 https://doi.org/10.1016/J.SCITOTENV.2023.163003.

Rubin, D.B., 1976. Inference and missing data. Biometrika 63 (3), 581–592. https://doi.org/10.1093/BIOMET/63.3.581.

Sadeghi, M., Nguyen, P., Hsu, K., Sorooshian, S., 2020. Improving near real-time precipitation estimation using a U-Net convolutional neural network and geographical information. Environ. Model. Software 134, 104856. https://doi.org/10.1016/J.ENVSOFT.2020.104856.

Sarma, T.A., 2013. Handbook of Cyanobacteria, 1st ed. CRC Press. https://doi.org/10.1201/b14316.

Scanlon, B.R., Jolly, I., Sophocleous, M., Zhang, L., 2007. Global impacts of conversions from natural to agricultural ecosystems on water resources: quantity versus quality. Water Resour. Res. 43 (3), 3437. https://doi.org/10.1029/2006WR005486.

Scutari, M., 2018. Dirichlet Bayesian network scores and the maximum relative entropy principle. Behaviormetrika 45 (2), 337–362. https://doi.org/10.1007/S41237-018-0048-X/FIGURES/4.

Scutari, M., 2019. Bayesian Network Models for Incomplete and Dynamic Data. http://arxiv.org/abs/1906.06513.

Simpson, I.M., Winston, R.J., Dorsey, J.D., 2023. Monitoring the effects of urban and forested land uses on runoff quality: implications for improved stormwater management. Sci. Total Environ. 862 https://doi.org/10.1016/J.SCITOTENV.2022.160827.

Tiyasha, Tung, T.M., Yaseen, Z.M., 2020. A survey on river water quality modelling using artificial intelligence models: 2000–2020. J. Hydrol. 585, 124670 https://doi.org/10.1016/J.JHYDROL.2020.124670.

van Buuren, S., 2016. Multiple imputation of discrete and continuous data by fully conditional specification, 16 (3), 219–242. https://doi.org/10.1177/0962280206074463.

Vigiak, O., Udías, A., Grizzetti, B., Zanni, M., Aloe, A., Weiss, F., Hristov, J., Bisselink, B., de Roo, A., Pistocchi, A., 2023. Recent regional changes in nutrient fluxes of European surface waters. Sci. Total Environ. 858 https://doi.org/10.1016/J.SCITOTENV.2022.160063.

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. Sci. Total Environ. 693, 133440 https://doi.org/10.1016/J.SCITOTENV.2019.07.246.

Wiering, M., Boezeman, D., Crabbé, A., 2020. The water framework directive and agricultural diuse pollution: fighting a running battle? Water (Switzerland) 12 (5), 1–12. https://doi.org/10.3390/w12051447.

Wright, S., 1921. Correlation and causation. J. Agric. Res. XX (7), 557–585, 1921.

Xiao, L., Liu, J., Ge, J., 2021. Dynamic game in agriculture and industry cross-sectoral water pollution governance in developing countries. Agric Water Manag 243, 106417. https://doi.org/10.1016/J.AGWAT.2020.106417.

Xu, J., Pan, J., Devlin, A.T., 2023. Variations in chlorophyll-a concentration in response to hydrodynamics in a flow-through lake: remote sensing and modeling studies. Ecol. Indic. 148, 110128 https://doi.org/10.1016/J.ECOLIND.2023.110128.

Zakwan, M., Wahid, A., Niazkar, M., Chatterjee, U. (Eds.), 2022. Water Resource Modeling and Computational Technologies. Elsevier.