Science of the Total Environment

Enhancing Environmental Data Imputation: A Physically-Constrained Machine Learning Framework

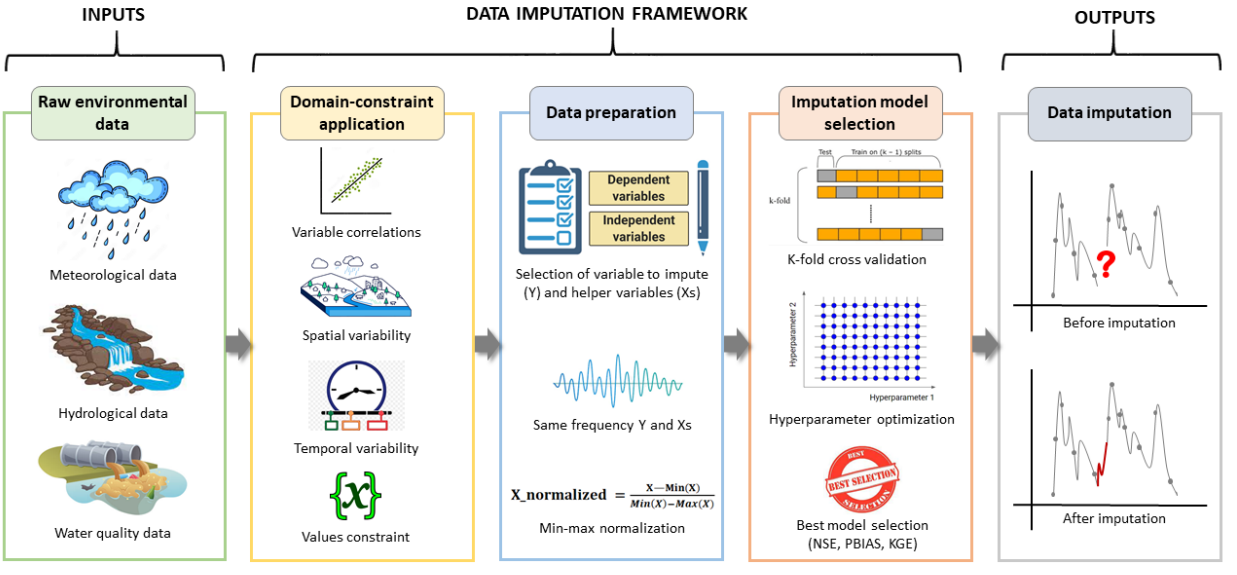
--Manuscript Draft--

Manuscript Number:	STOTEN-D-23-33253R2
Article Type:	Research Paper
Keywords:	environmental data, missing values, data imputation, machine learning, physical constraints.
Corresponding Author:	Angela Gorgoglione University of the Republic Uruguay URUGUAY
First Author:	Marcos Pastorini
Order of Authors:	Marcos Pastorini
	Rafael Rodríguez
	Lorena Etcheverry
	Alberto Castro
	Angela Gorgoglione
Abstract:	In water resources management, new computational capabilities have made it possible to develop integrated models to jointly analyze climatic conditions and water quantity/quality of the entire watershed system. Although the value of this integrated approach has been demonstrated so far, the limited availability of field data may hinder its applicability by causing high uncertainty in the model response. In this context, before collecting additional data, it is recommended first to recognize what improvement in model performance would occur if all available records could be well exploited. This work proposes a novel machine learning framework with physical constraints capable of successfully imputing a high percentage of missing data belonging to several environmental domains (meteorology, water quantity, water quality), yielding satisfactory results. In particular, the minimum NSE computed for meteorologic variables is 0.72. For hydrometric variables, NSE is always >0.97. More than 78% of the physical-water-quality variables is characterized by NSE>0.45, and more than 66% of the chemical-water quality variables reaches NSE>0.35. This work's results demonstrate the proposed framework's effectiveness as a data augmentation tool to improve the performance of integrated environmental modeling.
Response to Reviewers:	

Graphical Abstract

Enhancing Environmental Data Imputation: A Physically-Constrained Machine Learning Framework

Marcos Pastorini,Rafael Rodríguez,Lorena Etcheverry,Alberto Castro,Angela Gorgoglione



Highlights

Enhancing Environmental Data Imputation: A Physically-Constrained Machine Learning Framework

Marcos Pastorini,Rafael Rodríguez,Lorena Etcheverry,Alberto Castro,Angela Gorgoglione

- The novel framework successfully imputes multi-domain environmental data.
- It combines machine-learning algorithms with physical knowledge.
- It adequately works in case of a high percentage of missing values.
- It has a good generalization capability and can represent any scenario at a basin scale.
- Accurate data imputations will improve the performance of integrated models.

Enhancing Environmental Data Imputation: A Physically-Constrained Machine Learning Framework

Marcos Pastorini[a], Rafael Rodríguez[b], Lorena Etcheverry[a], Alberto Castro[a] and Angela Gorgoglione[b,*]

[a]Department of Computer Science, School of Engineering, Universidad de la República, Herreira y Reissig, 565, Montevideo, 11300, Uruguay

[b]Department of Fluid Mechanics and Environmental Engineering, School of Engineering, Universidad de la República, Herreira y Reissig, 565, Montevideo, 11300, Uruguay

ARTICLE INFO

Keywords:
environmental data
missing values
data imputation
machine learning
physical constraints

ABSTRACT

In water resources management, new computational capabilities have made it possible to develop integrated models to jointly analyze climatic conditions and water quantity/quality of the entire watershed system. Although the value of this integrated approach has been demonstrated so far, the limited availability of field data may hinder its applicability by causing high uncertainty in the model response. In this context, before collecting additional data, it is recommended first to recognize what improvement in model performance would occur if all available records could be well exploited. This work proposes a novel machine learning framework with physical constraints capable of successfully imputing a high percentage of missing data belonging to several environmental domains (meteorology, water quantity, water quality), yielding satisfactory results. In particular, the minimum NSE computed for meteorologic variables is 0.72. For hydrometric variables, NSE is always >0.97. More than 78% of the physical-water-quality variables is characterized by NSE>0.45, and more than 66% of the chemical-water quality variables reaches NSE>0.35. This work's results demonstrate the proposed framework's effectiveness as a data augmentation tool to improve the performance of integrated environmental modeling.

1. Introduction

1.1. Background and literature review

Over the past decade, there has been a notable increase in the utilization of integrated models for managing water quality concerns at the watershed scale (Freni et al., 2011). An integrated model is a specific model capable of simulating the interactions between multiple physical systems, such as the atmosphere, soil, and various water bodies (Freni and Mannina, 2012). These models are highly intricate and require substantial input data, parameters, and variables to maintain accuracy and reliability (Freni et al., 2009). It is of utmost importance to estimate missing data sequences within time series, as we need adequate

*Corresponding author
mpostorini@fing.edu.uy (M. Pastorini); rrodriguez@fing.edu.uy (R. Rodríguez); lorenae@fing.edu.uy (L. Etcheverry); acastro@fing.edu.uy (A. Castro); agorgoglione@fing.edu.uy (A. Gorgoglione)
ORCID(s): 0000-0003-0812-2419 (M. Pastorini); 0000-0003-3986-655X (R. Rodríguez); 0000-0001-8121-8076 (L. Etcheverry); 0000-0002-9174-398X (A. Castro); 0000-0002-2476-2339 (A. Gorgoglione)

environmental data to accurately represent the natural processes that occur and how the system responds at the catchment scale (Gorgoglione et al., 2020a).

Environmental time series may be incomplete due to technical issues with sensors or measurement instruments and data storage or transmission failures. Changes in the measurement site, data collectors, or instruments over time can also contribute to this (Chivers et al., 2020; Oriani et al., 2016; Sattari et al., 2017). To avoid spending a lot of time and money on collecting and analyzing further environmental records; it is essential first to understand how much the existing data can be improved (Gorgoglione et al., 2019,?). A methodology that can accurately fill in missing data from different but related environmental domains is necessary for this purpose.

Researchers have recently explored many approaches to minimize the missing data problem (Chen et al., 2021). Various techniques exist for managing missing data, from straightforward exclusion to more sophisticated imputation methods. To conduct statistical analysis, omitting all observations with missing values could work well if only a few observations contain unknown values (Bertsimas et al., 2018). Alternatively, it would introduce bias, and the information loss would often threaten the models' descriptive and predictive capabilities (White and Carlin, 2010). Furthermore, deleting observations would produce discontinuous time series, generating further difficulties in temporal data analysis.

When data is missing in a time series, data imputation can estimate the missing values and maintain the length of the series. One standard method is statistical analysis, using mean, median, or mode to fill in missing data (Kabir et al., 2020). However, this technique can result in flat imputed values (Chen et al., 2021). In the environmental domain, observations from neighboring monitoring stations can also replace missing data. However, this may only sometimes be reliable due to weak correlations at longer distances (Blenkinsop et al., 2017). Distance-based weighted interpolation techniques have been used for missing meteorological data. Still, they may not account for the non-linear spatiotemporal relationships that describe most environmental variables, especially if the variables under study are water-quality related.

With this purpose, many multivariate methods have been proposed, including hot-deck imputation, expectation maximization, predictive mean matching, least squares regression, support vector regression, gradient boosting, nearest neighbor techniques, decision tree techniques, and artificial neural networks (Andridge and Little, 2010; Bertsimas et al., 2018; Bø et al., 2004; Dempster et al., 1977; Gill et al., 2007; Honaker et al., 2009; Körner et al., 2018; Templ et al., 2011; Troyanskaya et al., 2001; Wang et al., 2006). In

addition, supervised machine learning techniques can effectively represent non-linear relationships between variables measured at different spatially distributed stations (Chivers et al., 2020). However, many machine learning methods fail to consider data's temporal variability, which hampers their accuracy in imputing variables that exhibit predictable temporal patterns. To effectively impute water quality data at the catchment scale, it is crucial to integrate environmental data from interconnected domains such as meteorology and hydrology. This integration enables a comprehensive understanding of water quality's dynamic nature and enhances imputation techniques' accuracy.

1.2. Related work

Various machine learning techniques have been used to address missing data in environmental data sets, including in the fields of meteorology, hydrology, and water quality (Chandra et al., 2021; Chrobak et al., 2022; Tencaliec et al., 2015). Researchers have recently focused on addressing data imputation in the water-quality domain. For example, Chen et al. (2021) developed a new TrAdaBoostLSTM framework combining deep learning and transfer learning to impute large-scale consecutive missing data. The framework also employs the dynamic time-warping method to identify the source domain with complete data most similar to the target domain with incomplete data. This approach imputes the dissolved oxygen concentration data from ten monitoring stations in the Qiantang River basin in China. Tabari and Hosseinzadeh Talaee (2015) evaluated the efficiency of the multilayer perceptron (MLP) and radial basis function (RBF) networks for reconstructing the missing values of thirteen water quality variables at five monitoring stations in the Maroon River basin, Iran. They concluded that the MLP outperforms the RBF networks for this purpose. Bi et al. (2022) proposed a method based on generative adversarial networks applied for the first time to impute water quality data (water temperature, pH value, total nitrogen, and dissolved oxygen). Such time series were collected at one monitoring station in China and characterized by a maximum data missing rate of 30%.

Although several methods are available for imputing missing data, only a few effectively handle a high percentage of missing values. Aguilera et al. (2020) conducted a study to compare the performance of spatiotemporal kriging (STK), random forest (RF) algorithm, and multiple imputations by chained equations through predictive mean matching (PMM) in imputing daily precipitation data from 112 rain gauges in southwestern Spain. The study tested different percentages of missing data (ranging from 64% to over 90%) and missing patterns. The results showed that STK performed better than PMM and RF in simulating the

precipitation distribution under missing chronological patterns, although it had a higher computing cost. Meanwhile, RF was an efficient alternative for imputing daily precipitation data, especially in random missing patterns.

Ratolojanahary et al. (2019) combined multivariate imputations by chained equations (MICE) with different machine learning models (RF, boosted regression trees (BRT), K-nearest neighbors (KNN), and support vector regression (SVR)) to tackle multiple correlations between a high amount of water quality variables (257) and a high rate of missing data (more than 80%). The research findings showed that combining MICE with SVR, RF, KNN, and BRT outperforms the original MICE alone. Moreover, MICE-SVR represents a good trade-off regarding computing time and performance. Jones et al. (2014) evaluated the performance of MICE to impute the values of six chemicals in community water systems. The technique was applied in a simulated environment using data from the Atrazine Monitoring Programme in five Midwestern US states, where 65-92% of the observations were suppressed. The authors found multiple imputations to be an effective method to fill in water-quality data.

A recent study by Zhang and Thorburn (2021) developed a deep neural network architecture (Dual-SSIM) for hydrologic (water level, discharge) and water quality (water temperature, conductivity, turbidity, nitrate) data imputation. Experimental results demonstrated that Dual-SSIM outperformed other benchmarks such as Expectation Maximization, K-nearest neighbor, sequence-to-sequence architecture with global attention mechanisms (SSIM), recurrent neural network-based method (BRITS), and Multi-directional Recurrent Neural Networks. This method was successfully applied in Iowa River (USA) and Russel River (Australia) (Zhang and Thorburn, 2022).

While many studies have developed imputation methods or compared various algorithms for specific environmental variables, it remains a challenge to determine the most effective technique for variables belonging to different environmental domains and under various circumstances.

1.3. Objective and contributions

This study presents a novel machine-learning framework that incorporates physical constraints to support various imputation algorithms. Our framework offers two key advancements: *i*) It can effectively and simultaneously impute many variables from different environmental fields, including meteorology, hydrology, and water quality (physical and chemical); *ii*) It adopts a physically-constrained approach,

enabling the integration of data-driven algorithms with the temporal and spatial variability of the variables and their correlations.

Compared to previous methods, our approach offers several advantages. Firstly, it can successfully handle a high percentage of missing values. Secondly, it takes a multivariate approach by considering many variables simultaneously. Lastly, it leverages a wide range of statistical and machine-learning techniques.

What distinguishes our approach from previous studies is its systematic strategy to addressing missing data in meteorological, hydrometric, and water quality variables at the catchment scale. This approach can be applied to any environmental dataset, improving water quality simulation and prediction. The accurate imputation of missing environmental data not only refines our understanding of specific environmental dynamics but holds broader implications for informed environmental management and decision-making, enabling policymakers to develop targeted strategies based on a more comprehensive and reliable dataset. It is important to note that this work builds upon the study conducted by Rodríguez et al. (2021), which evaluated the performance of various statistical and machine learning algorithms for imputing water quality data with a high percentage of missing values at six monitoring stations in the Santa Lucía Chico river, Uruguay.

Our data imputation framework is expected to significantly improve environmental model accuracy and enable a more comprehensive understanding of environmental dynamics. This not only accelerates scientific advancements but also directly aids policymakers in developing precise, data-driven policies for real-world challenges.

2. Materials

2.1. Study area

The study area is the Santa Lucía Chico (SLC) watershed, located between S33°42′ - S34°50′ and W55°0′ - W57°6′ (Figure 1). It is one of the most critical watersheds in Uruguay (South America) since it is the country's primary source of drinking water and also supports numerous agricultural and industrial activities (Navas et al., 2019; Vilaseca et al., 2023). The primary national drinking water reservoir, Paso Severino, is located in this watershed. This catchment was chosen not only for its strategic importance but also because it is a mixed lotic and lentic system (Rodríguez et al., 2021). We found it interesting to test our framework with data recorded at sites characterized by different hydraulic and hydrologic conditions

to analyze its generalization capabilities. SLC catchment has a surface equal to 2570 km^2 and a perimeter of 300 km. Its elevation ranges between 177 m a.s.l in the northeast area, 25 m a.s.l at the Paso Severino reservoir, and 3 m a.s.l at the main channel outlet ([dataset] MGAP, 2020). The average slope of the basin is 2.68%, and the length of the main channel is 128.7 km. The catchment is in a temperate climate zone characterized by four seasons (Gorgoglione et al., 2020b). Average annual temperatures can vary between 3 °C (in winter) and 30 °C (in summer). Annual precipitation can vary between 1000 mm and 1500 mm (INUMET, 2020).

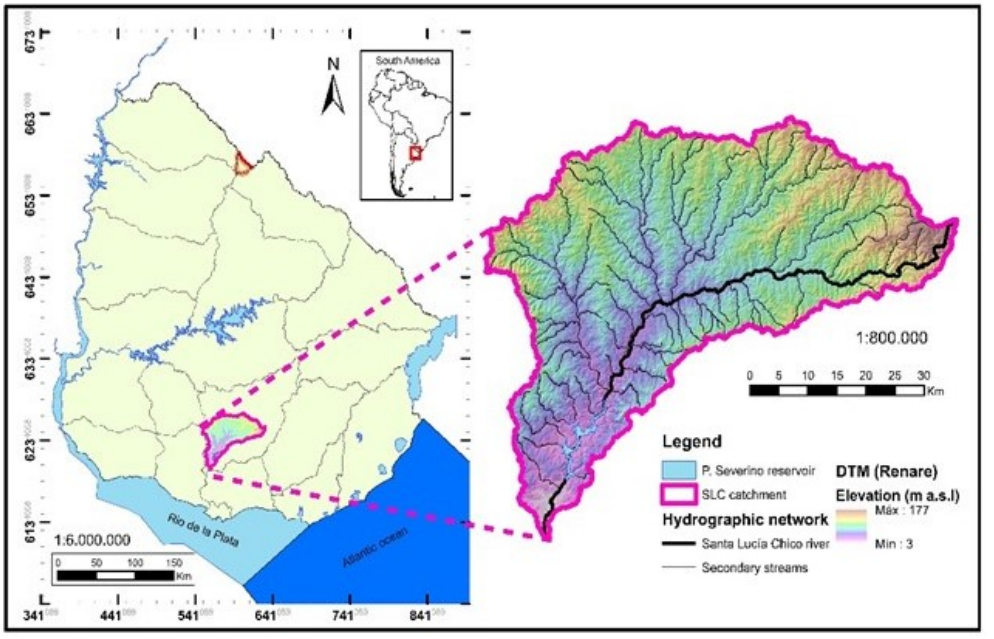


Figure 1: Location of the Santa Lucía Chico watershed, with its hydrographic network and DTM elevation. Coordinate system WGS 84/UTM 21s.

2.2. Data description

The dataset selected for this study includes the following three groups of variables: **i) meteorological variables:** precipitation (P) [mm], evapotranspiration (ET) [mm], air temperature (average (T_{ave}), maximum (T_{max}), and minimum (T_{min})) [°C], solar radiation (SR) [cal/cm^2/d], heliophany (sunshine hours) (Hel) [!htb], average relative humidity (RH) [%], and wind speed (WS) [2m/km/d]; **ii) hydrometric variables:** streamflow (Q) [m^3/s], water level (h) [m]; **iii) physical and chemical water-quality variables:** water temperature (WT) [°C], conductivity ($Cond$) [μS/cm], nitrite (NO_3^-) [mg/L], nitrate (NO_2^-) [mg/L], ammonium (NH_4^+) [mg/L], total nitrogen (TN) [mg/L], dissolved oxygen (DO) [mg/L], potential of hydrogen (pH) [NA],

turbidity (*Turb*) [NTU], phosphate (PO_4^{3-}) [mg/L], total phosphorus (*TP*) [μg *P*/L], chlorophyll-a (*Chl-a*) [μg/L], biochemical oxygen demand (*BOD*) [mg/L], total suspended solids (*TSS*) [mg/L], total solids (*TS*) [mg/L]. The dataset corresponds to the period 2014-2020, except for the variables PO_4^{3-}, *TSS*, and *TS*, which were monitored from 2018 and were discarded from the imputation process.

The National Institute of Agricultural Research (INIA) and the Uruguayan Institute of Meteorology (INUMET) collected the meteorological dataset. The following variables (*ET*, *RH*, T_{ave}, T_{max}, T_{min}, and *WS*) were measured at "Las Brujas" meteorological station once a day. The dataset corresponds to the period 1/8/2014 - 31/12/2020, and it is freely downloadable from the data bank of the Agroclimate and Information Systems Unit (INIA, 2020). Precipitation was collected from nine INUMET conventional rain gauges and a meteorological station with a daily frequency from 1/8/2014 - 29/6/2020.

This study also analyzed a hydrometric dataset collected by the Uruguayan National Water Board (DINAGUA) from August 1, 2014 to June 30, 2020. The water level (*h*) was measured three times a day at hydrometric stations in Florida to determine streamflow ((*Q*)). The National Board for Quality and Environment Assessment (DINACEA) gathered a water-quality dataset from 2014 to 2020. This data is free and available to the public through the National Environmental Observatory (OAN) ([dataset] DINACEA, 2020). The dataset was collected at six monitoring stations along the SLC River. Three of these stations (SLC01, SLC02, PS01=SLC03) are located upstream of the Paso Severino reservoir, while the remaining ones (PS03, PS04, and PS02) are in the reservoir. To give a visual representation of the stations and measurement points, refer to Figure 2.

A summary of the datasets used in this study, along with the percentage of missing values detected for each variable, is reported in the Supplementary Information (A.1).

3. Methods

3.1. Methodology conceptualization

We designed and implemented a novel missing data imputation framework in this study. The methodology comprises four main phases: in *Phase 1*, we apply physical constraints to our framework that considers variables' temporal and spatial variability, correlations, and range of variation. Each of these constraints is implemented as an artificial variable (column) that is added to the original dataset, creating, in this way, a new dataset variant. *Phase 2* involves selecting helper variables for data preparation to aid imputation. These

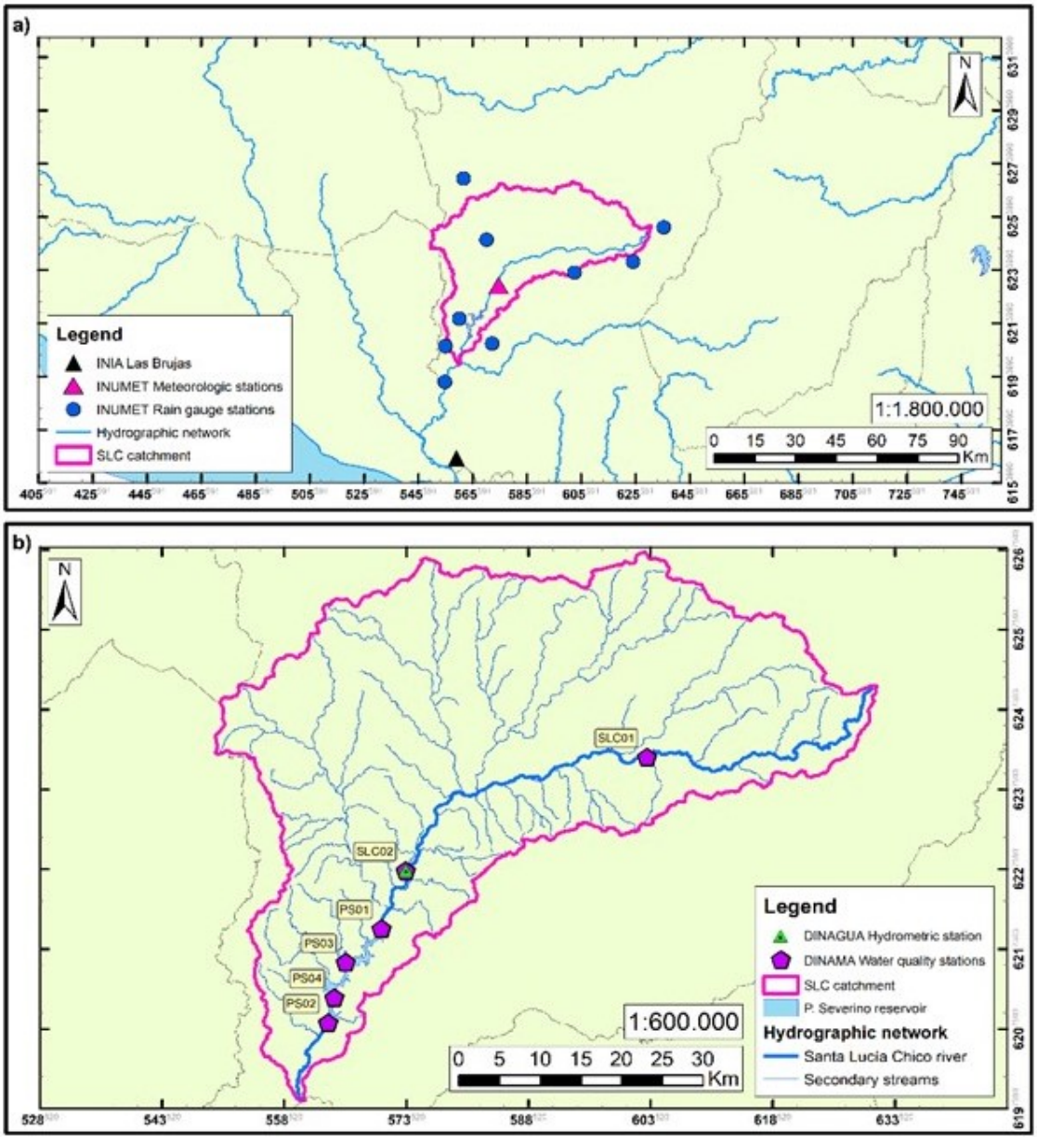


Figure 2: Location of monitoring stations. Coordinate system WGS 84/UTM 21s.

variables will be used to train imputation models. In *Phase 3*, we train and test these models to determine the best one for data imputation. Finally, we generate the imputed dataset in *Phase 4*. We follow an iterative process, starting with variables with the least missing data. These variables then act as helper variables to complete missing values in other time series.

Figure 3 depicts the proposed methodology, while each phase is thoroughly described in the following sections. Our framework was implemented using Python on a desktop computer (Ubuntu Operating System, 16 GB of RAM, and Intel i3 Processor), and the front end was developed using the Streamlit library.

3.2. Definition and application of domain constraints

As mentioned, Phase 1 involves applying physical domain constraints to the framework, considering variables' temporal and spatial variability, correlations, and range of variation. We analyzed the variables' temporal patterns, spatial variations, and correlations to incorporate relevant constraints. Additionally, we considered each variable's permissible range of values to prevent unrealistic imputations. Doing so establishes a solid foundation for accurate and reliable imputations in subsequent phases. The following subsections will provide detailed definitions and procedures for each domain restriction.

3.2.1. Variable correlation

The machine-learning framework considers the correlation between variables using Pearson, Spearman, and Kendall correlation matrices with a $> |0.5|$ threshold. All three matrices produced similar results, and Figure 4 shows the Spearman correlation matrix with each line representing a variable at a monitoring station. Similar variables were grouped together and reported only once to avoid confusion, while Pearson and Kendall's matrices can be found in the Supplementary Information (A.2).

We will now focus on the strongest correlations we have identified, but a full list can be found in the Supplementary Information (A.2) for your reference. The positive relationship between *WT* and the meteorological variables, particularly *SR*, air temperature, and *Hel* is clear. A strong inverse correlation exists between *WT* and *DO*, supported by warm water holding less *DO* than cold water. In winter and early spring, when the *WT* is low, the *DO* concentration is high; in summer and early fall, when the water temperature rises, the *DO* concentration is often lower (Gorgoglione et al., 2020b; Rodríguez et al., 2021). Based on such relationships, it is easy to understand the inverse correlation between *DO* and the meteorological variables, especially *SR*, air temperature, and *Hel*.

Furthermore, *Cond* is highly influenced by *WT* and *Turb*. An increase in *WT* determines a more significant amount of ions due to molecule dissociation and an increase in ionic mobility. Since *Cond* depends on such factors, an increase in *WT* causes an increase in *Cond* Hayashi (2004). Moreover, *Cond* represents the ability of a liquid to conduct an electric charge, which depends on dissolved ion concentration, usually

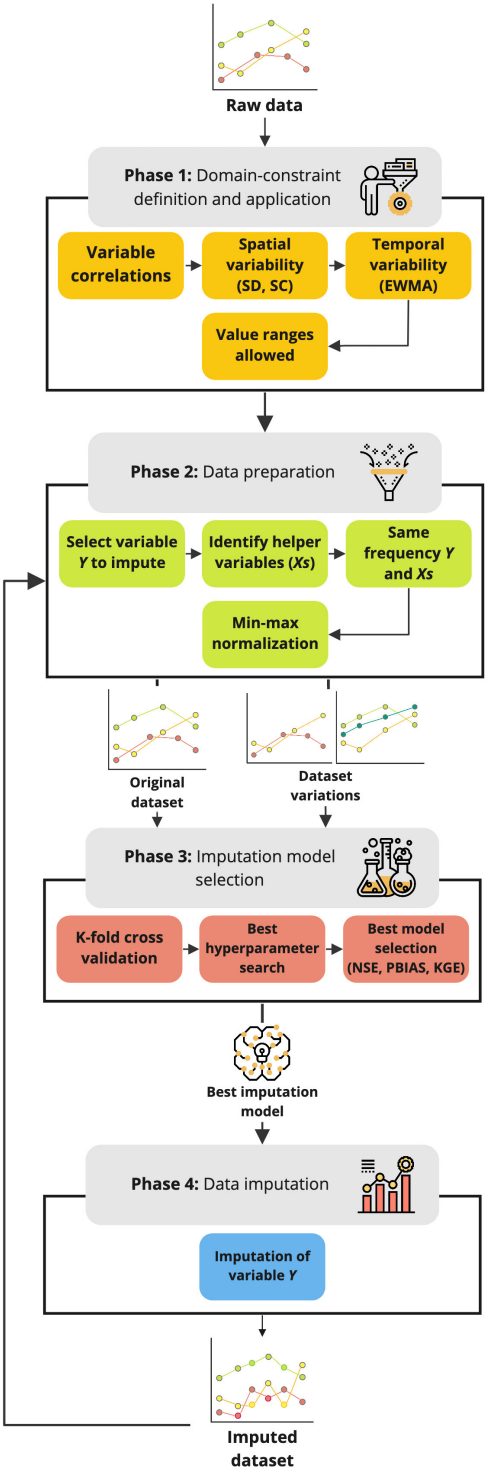


Figure 3: Conceptualization of the proposed methodology.

measured as total dissolved solids (Bakhtiar Jemily et al., 2019). Since the latter is highly correlated to *Turb*, it is easy to justify the strong relationship between *Cond* and *WT* and, consequently, between *WT*

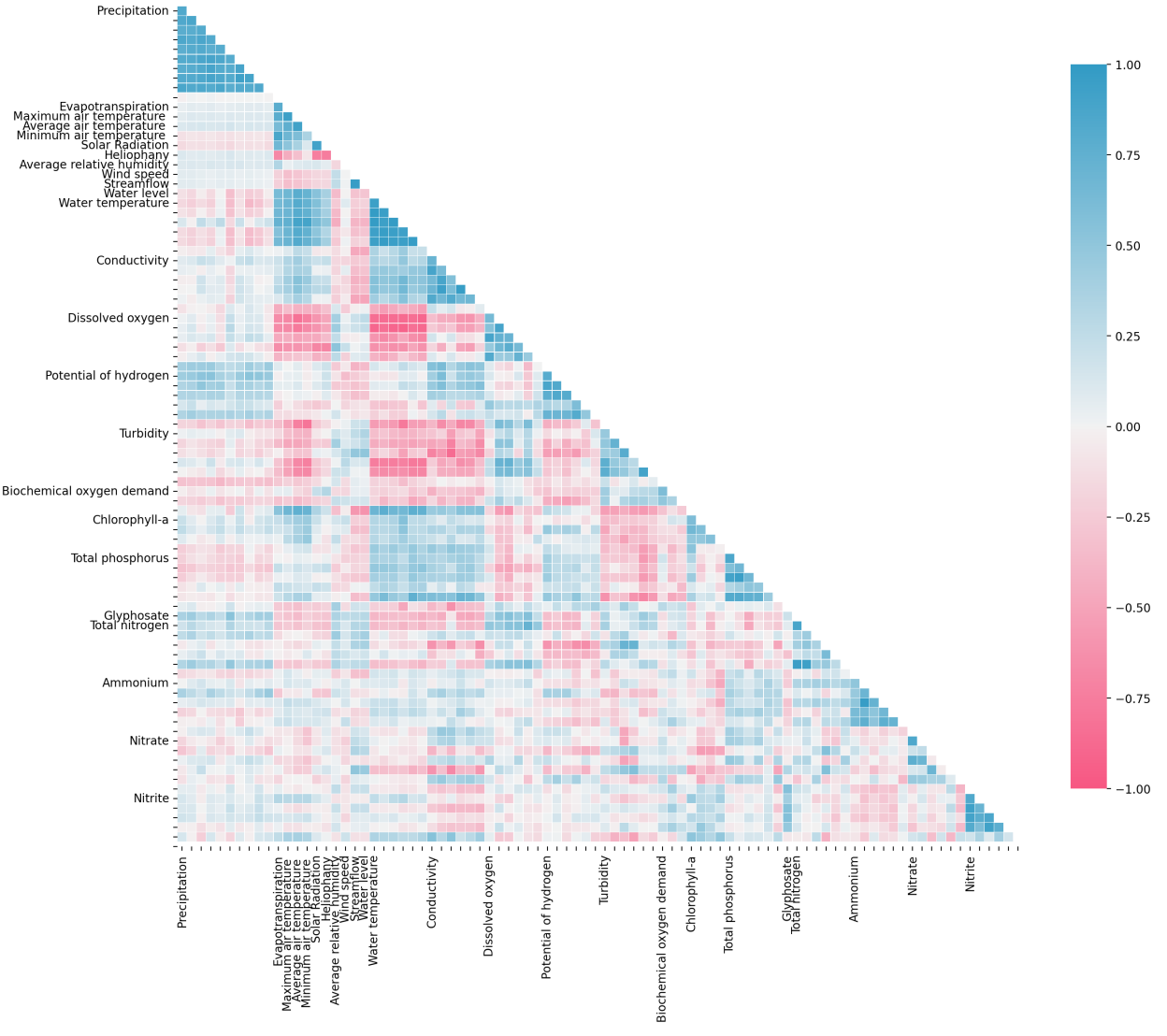


Figure 4: Spearman correlation matrix.

and *Turb*. The latter is also supported by the fact that suspended particles in water bodies, whose proxy is *Turb*, absorb heat from *SR* more efficiently than water. Such heat is then transferred from the particles to water molecules, raising, in this way, the surrounding *WT* (Paaijmans et al., 2008). Based on the above, the negative correlations between *Turb* and the meteorological variables, particularly air temperature, and between *DO* and *Cond* are simple to understand.

The variables *TP* and *WT* are also directly correlated. The increase in *WT* improves the microorganism activity and phosphorus diffusion from the interstitial water to the overlying water, thus affecting phosphorus release (Cheng et al., 2020). Such a relationship explains the robust direct correlation between *TP* and *Cond*.

TP also resulted highly correlated to *Turb*, which is especially true in watersheds characterized by a high proportion of agricultural land like the one under study (Villa et al., 2019). This phenomenon is explained by the fact that most phosphorus export occurs from agricultural areas due to solids off-site movement, which carries the phosphorus to water bodies.

We also considered the correlations between *Chl-a* with *WT*, *Cond*, and *Turb*, respectively. Growth of cyanobacteria in freshwater containing only *Chl-a* is generally favored at higher temperatures, with well-defined thermal optima for growth at temperatures ranging from 20 to 30 °C (Haakonsson et al., 2017). The correlation *Chl-a-Cond* depends on the cyanobacteria species. Haakonsson et al. (2020) found a negative correlation between these two variables at Punta del Tigre in the Río de la Plata Estuary, Uruguay, since high salinity limits or inhibits cyanobacterial growth.

The positive correlation *Chl-a-Turb* depends on cyanobacteria being part of the suspended particles that contribute to *Turb*. Crisci et al. (2017) found that *WT*, *Cond*, and *Turb* were among the most relevant phytoplankton biomass predictors at Laguna del Sauce, Uruguay. The results obtained from the three correlation matrices were complemented by other variable correlations presented in the scientific literature.

Figure 5 displays a dependency tree that summarizes the correlations between variables used in our framework. The tree highlights variables that were imputed with the help of one, two, or three other variables (helper variables) are represented in yellow, orange, and red, respectively. Meteorological variables used as helpers are depicted in green since their imputation doesn't depend on any other imputation. Table 9 provides the scientific literature that supports each correlation reported in Figure 5.

3.2.2. Spatial variability

We evaluated and incorporated spatial dependencies (SD) and spatial corrections (SC) between various monitoring stations to account for spatial variability. SD considers the spatial placement of monitoring sites, including upstream and downstream locations. For example, knowing that Q is influenced by P (Section 3.2.1), the imputation of Q-time series will be helped by P-times series monitored at those upstream Florida stations, where Q was monitored. The SD implemented in the framework is depicted in Figure 6. Additionally, the SC represents the weight given to helper variables based on how far they are from the

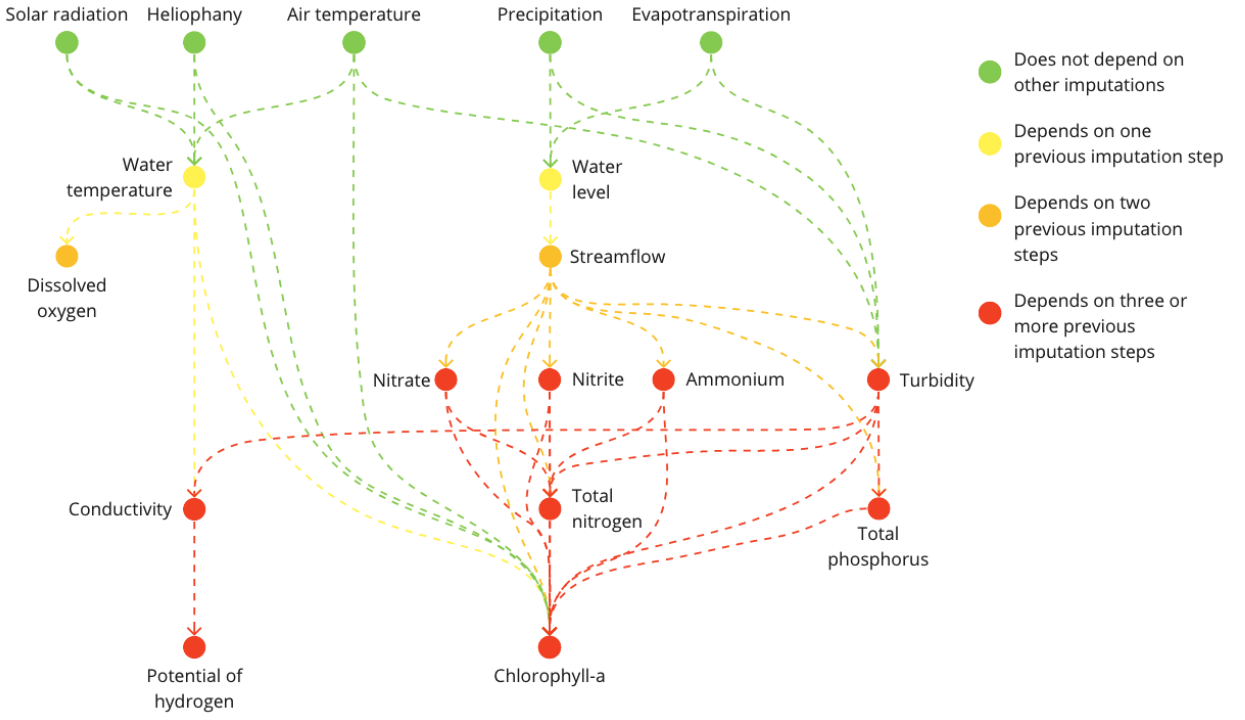


Figure 5: Variable-dependency tree for data imputation.

monitoring stations of the imputation variable. The closer the measured values are to the imputation location, the more influence they have, according to the *Inverse Distance Weighted* (IDW) method.

IDW assumes that the influence of each measured point decreases as the distance increases:

$$Y_m = \frac{\sum_{i=1}^{n} Y_i d_{mi}^{-k}}{\sum_{i=1}^{n} d_{mi}^{-k}} \tag{1}$$

where Y_m is the observation at station m, n is the number of stations, Y_i is the observation at station i, d_{mi} is the distance between m and i, k is the exponent that generally ranges between 1 and 6. In this study, $k = 2$ was assumed. Two separate dataset variants were created within the framework to complete the imputation process: Dataset SD and Dataset SC.

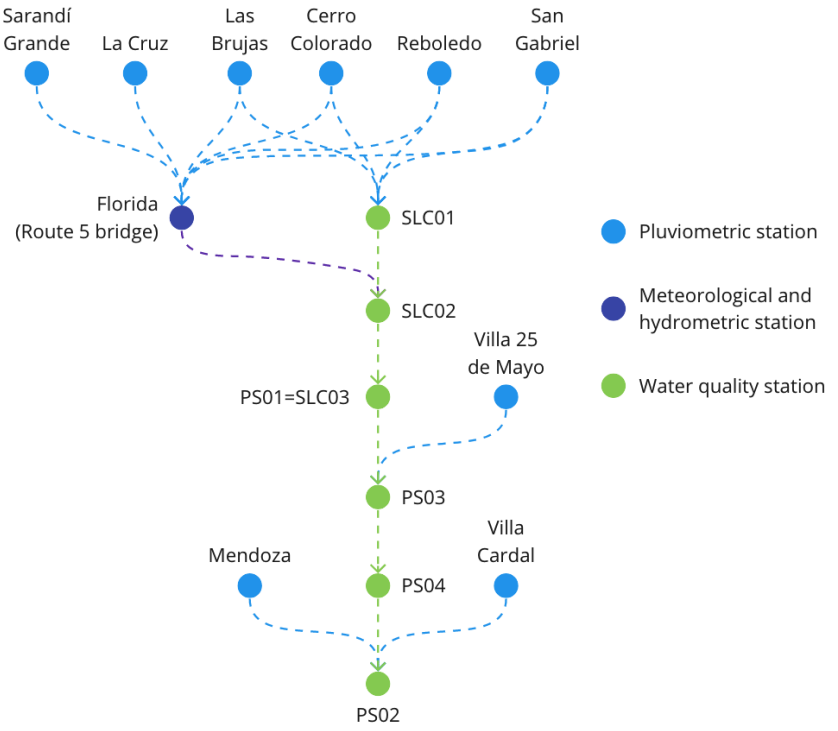


Figure 6: Spatial dependency tree.

3.2.3. Temporal variability and value constraints

We used the *Exponentially Weighted Moving Average* (EWMA) to account for the temporal variability in the studied variable. This method gives more weight to recent observations and less weight to older ones, based on an exponential decay function:

$$\begin{cases} \text{EWMA}(Y_n) = \dfrac{\sum_{i=0}^{t}(1-\alpha)^i Y_{n-1-i}}{\sum_{i=0}^{t}(1-\alpha)^i} & \alpha = \dfrac{2}{t+1} \end{cases} \tag{2}$$

where Y is a time series, n is the number of observations, α is the weight assigned, and t is the temporal time window selected based on the variable to impute: two months for water-quality variables, one week for hydrometric variables, and one day for climatic variables.

A new dataset variant called "Dataset EWMA" was created by adding the EWMA artificial variable as a new feature. In addition, specific ranges of variation were imposed for all variables in the models. The output variables can range from 0 to $+\infty$, while pH can range from 0 to 14.

3.3. Data preparation

Before conducting any analysis, it was important to address the challenge of dealing with variable names that were not standardized, different units of measurement, varying orders of magnitude, and sampling frequencies. To address this, a group of helper variables (Xs) was selected along with the variable Y that needed to be imputed based on the domain constraints outlined in Section 3.2. If a helper variable had less than 50% missing values, it was temporarily imputed using linear regression. However, if the percentage of missing values was greater than 50%, the variable was not considered a helper to avoid introducing noise. If a helper variable $X \in Xs$ had a different monitoring frequency than Y, it was substituted with its maximum (X_{max}), mean (X_{mean}), and minimum (X_{min}) to match the frequency of Y. Furthermore, all variables were reduced to a monthly frequency, which was the sparser one that characterized the water-quality variables. A min-max normalization approach was implemented to equalize the importance of each variable and manage their different measurement units. Finally, the data were tabulated, resulting in a dataset where each data point (row) included time information (in the form of new artificial variables, i.e., EWMA column). In this way, data points were independent from each other.

Moreover, the dataset variants generated from the domain-constraint application were created and added to Xs. It is important to remark that the framework trained and tested all the imputation models using the original dataset (Original dataset), all the dataset variants (Dataset EWMA, Dataset SD, Dataset SC), and their combinations (Dataset EWMA+SD, Dataset EWMA+SC, Dataset SD+SC, Dataset EWMA+SD+SC). The output of the framework is the best dataset-model pair. It is selected based on a comprehensive evaluation across these diverse datasets and the different models.

3.4. Imputation models

We examined various models to determine the most effective ones for imputing different variables from various environmental domains, as outlined in Wolpert and Macready (1997). Since many of the variables, especially those related to water quality, had limited data available, it was crucial that the chosen methods had strong predictive capabilities with a smaller amount of data.

Imputation methods can be grouped into two categories (Durbin and Koopman, 2012): i) univariate: algorithms that only take into account the values of the imputing variable; ii) multivariate: methods that, besides the values of the imputing variable, also consider other variables' data-points as input. Table 1 lists

Table 1
Imputation algorithms taxonomy.

	Single model	Multi-model
Univariate	Inverse Distance Weighting (IDW)	
Multivariate	Ridge Regressor (RR) TheilSen Regressor (TR) Huber Regressor (HR) Bayesian Ridge Regressor (BRR) Support Vector Regressor (SVR) K-nearest neighbors Regressor (KNNR)	Random Forest Regressor (RFR) Multivariate Imputation by Chained Equations (MICE)

the imputation methods implemented in the framework, while a brief description of each method is given in the following section.

3.4.1. Univariate imputation methods

The ***Inverse Distance Weighting*** **(IDW)** interpolation model assumes that closer objects are more similar than those farther apart. It estimates unmeasured values using observed values from nearby locations, with greater impact from closer locations. Weights decrease as the distance from the imputation location increases (Fortin and Dale, 2005). We implemented the model using *numpy* (Harris et al., 2020) and *pandas* (McKinney, 2010) libraries.

3.4.2. Multivariate imputation methods

The methods in this category are based on a set of simple regression models and machine-learning-aided regression models. In all the following cases, our implementation is based on Python's *scikit-learn* library (Pedregosa et al., 2012).

The ***Random Forest Regressor*** **(RFR)** is a machine-learning method that utilizes an ensemble of decision trees (Breiman, 2001). Decision trees are structures that divide input-feature space into smaller subspaces (Stockman et al., 2019). The RF method trains each decision tree on a different set of data points obtained by bootstrapping, and each tree may include a different subset of randomly chosen input features. The RF method's output is obtained by aggregating the outcomes of all decision trees, which is done by considering the mean for regression problems (Mital et al., 2020). We utilized Python's *scikit-learn* Extremely Randomized Trees Regressor. This regressor differs from the RFR in the node division decision method, where each division is made randomly instead of searching for the optimal cut. This change reduces training time without affecting prediction power (Geurts et al., 2006).

The ***Ridge Regressor*** **(RR)** is a tool for estimating regression coefficients for high dimensional data where the dataset contains correlated features (Hoerl and Kennard, 2000). This is often the case for environmental data. RR can be used to get stable parameter estimates when standard multiple regression methodologies fail. RR coefficients can be efficiently calculated by computing an orthogonal transformation of the high-dimensional data (Cule and De Iorio, 2012).

The ***TheilSen Regressor*** **(TR)** trains a regression model based on data statistics instead of single points to make it robust to outliers (Dang et al.), while the ***Huber Regressor*** **(HR)** trains a regression model optimizing the squared loss or the absolute loss depending on the samples used. This approach allows the model not to be heavily influenced by outliers while still taking their effect into consideration (Owen, 2006).

The ***Bayesian Ridge Regressor*** **(BRR)** consents to a natural mechanism to survive poorly distributed or insufficient data by formulating linear regression employing probability distributors rather than point estimates. The output is assumed to be drawn from a probability distribution rather than estimated as a single value (Tipping, 2001).

The ***Support Vector Regressor*** **(SVR)** uses Support vector machines. These algorithms look for a hyperplane or a set of them in data, which is non-linearly transformed into a higher dimensional space through kernel methods (Suykens and Vandewalle, 1999). The hyperplane and boundary layers minimize an error function for regression applications to estimate equation coefficients (Chivers et al., 2020).

The ***K-nearest neighbors Regressor*** **(KNNR)** uses the nearest neighbors algorithm: a non-parametric technique. In the feature space, some nearest neighbors are weighted based on a distance function chosen by the user (Euclidean distance is the most commonly used). The output is the average of the k nearest neighbors (Kramer, 2013).

Finally, we implemented the ***Multivariate Imputation by Chained Equations*** **(MICE)** based on each of the previous multivariate imputation models. It operates under the hypothesis that given the variables used in the imputation process, the missing data are missing at random (MAR), assuming that missing value probability depends exclusively on recorded values (Graham, 2009). In other words, after checking for all available data (i.e., the variables included in the imputation model), any remaining missing information is entirely random (Azur et al., 2011). With MICE, a base regression model is selected and then used for imputing each variable with missing values. Here, each variable with missing data is iteratively modeled based on the other variables in the dataset. The use of MICE generates a model variant hereafter called

Table 2
Summary of the dataset and model variants.

Variant		Description
Dataset and model	Original	Model and data without any variant
Dataset	SC	Spatial correlations (Equation 1)
	SD	Spatial dependency (Figure 6)
	EWMA	Temporal variability (Figure 2)
Model	MICE	Base model retrained with MICE

"MICE." Table 2 summarizes all the datasets and model variants described in this section and implemented in our framework.

3.5. Model cross-validation and data imputation

The framework implemented cross-validation to ensure the best model for each variable by splitting the input dataset into approximately equal-sized groups (folds). The first fold was used as the validation set, and the rest as the training set. This process was repeated k times (we chose 10-fold cross-validation for the input datasets), and the average loss-function values (*Nash-Sutcliffe efficiency* (NSE), *percent bias* (PBIAS), and *Kling-Gupta efficiency* (KGE)) were calculated. If the input dataset had less than 100 observations, we ran repeated k-fold cross-validation, with $k = \max(N/10, 2)$ (where N is the number of data points), n times, with $n = 10/k$, randomly selecting folds during any iteration. This ensured that the number of times each performance metric was measured was equal to the classical k-fold cross-validation.

To ensure accuracy and reliability, we conducted a validation process with and without repetitions and tuned hyperparameters using the Python library *Optuna*, which is open-source (Akiba et al., 2019). Our objective function was to select the best model for each variable based on the highest NSE. This model imputes the selected variable Y during Phase 2. If there were more variables to impute, the imputed Y could be used as a helper variable to complete those time series. Once all variables were imputed, we obtained the final complete dataset.

3.6. Model performance evaluation

To evaluate the performance of the imputation models, we calculated and compared NSE, PBIAS, and KGE:

$$\mathrm{NSE} = 1 - \frac{\sum_{i=1}^{n}(y_i^o - y_i^m)^2}{\sum_{i=1}^{n}(y_i^o - \bar{y^o})^2} \tag{3}$$

$$\mathrm{PBIAS} = 100 \cdot \frac{\sum_{i=1}^{n} y_i^o - y_i^m}{\sum_{i=1}^{n} y_i^o} \tag{4}$$

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \tag{5}$$

where y_i^o is the i^{th} observed value, y_i^m is the corresponding modeled value (imputed), $\bar{y^o}$ is the mean of observed values, and n is the size of the testing dataset. Being μ_m, δ_m and μ_o, δ_o mean and standard deviation (the first two statistical moments) of y^m and y^o, respectively; r is the linear correlation between observations and imputations, α is a measure of the flow variability error ($\alpha = \mu_m/\mu_o$), and β is a bias term ($\beta = \delta_m/\delta_o$).

The NSE is a normalized statistical method that defines the relative magnitude of the residual variance of a model compared to the variance of measured data (Nash and Sutcliffe, 1970). It ranges between -∞ and 1. If NSE=1, imputed values perfectly reproduce the observed ones. If NSE=0, imputed values are only as good as the observation mean. If NSE<0, the mean observation value is a better predictor than imputed values. Therefore, higher NSE values are preferable since they imply a more accurate imputation model.

PBIAS measures the average tendency of the imputed data to be smaller or larger than their observed counterparts (Moriasi et al., 2007). PBIAS=0 is the optimal value, with low-magnitude values representing accurate model imputation. Negative values characterize model overestimation bias, while positive values represent model underestimation bias.

Finally, KGE represents the Euclidean distance computed using the coordinates of r, α, and β (Gupta et al., 2009). As well as NSE, it ranges between -∞ and 1. However, there are no well-defined KGE thresholds that outline a "good" model as for NSE (Knoben et al., 2019; Rodríguez et al., 2021). The benchmark associated with NSE estimates (i.e., NSE=0) occurs when the estimate of KGE=$1 - \sqrt{2}$, i.e., when the estimate of KGE=-0.41 (Knoben et al., 2019).

Table 3
Performance ratings defined for each evaluation metric.

Performance rating	Hydrometric and climatic variables	Physical water quality variables	Chemical water quality variables
NSE			
Very good	NSE > 0.80	NSE > 0.80	NSE > 0.65
Good	0.70 <NSE≤0.80	0.70 < NSE≤0.80	0.50 <NSE≤ 0.65
Satisfactory	0.50 <NSE≤0.70	0.45 <NSE≤0.70	0.35 <NSE≤ 0.50
Unsatisfactory	NSE ≤ 0.50	NSE ≤ 0.45	NSE ≤ 0.35
PBIAS			
Very good	|PBIAS|< 5	|PBIAS|< 10	|PBIAS|< 15
Good	5 ≤|PBIAS|< 10	10 ≤|PBIAS|< 15	15 ≤|PBIAS|< 20
Satisfactory	10 ≤|PBIAS|< 15	15 ≤|PBIAS|< 20	20 ≤|PBIAS|< 30
Unsatisfactory	|PBIAS|≥ 15	|PBIAS|≥ 20	|PBIAS|≥ 30
KGE			
Satisfactory/Good	KGE ≥ -0.41	KGE≥ -0.41	KGE≥ -0.41
Unsatisfacory	KGE < -0.41	KGE < -0.41	KGE < -0.41

Table 3 summarizes the performance ratings defined for each evaluation metric (NSE, PBIAS, KGE). These metrics are derived from previously published studies (Chen et al., 2017; Moriasi et al., 2015; Rodríguez et al., 2021). NSE was chosen as the objective function because of its strict standards for determining a good fit. PBIAS and KGE were also calculated to validate the accuracy of each model used in the study.

4. Results and discussion

4.1. Imputation results

To evaluate the performance of the implemented imputation models within the framework and determine the most suitable model for each variable, a 10-fold cross-validation approach was employed. In cases where a time series had fewer than 100 records, repeated 10-fold cross-validation was used. The best model selection for each variable was based on the highest Nash-Sutcliffe Efficiency (NSE) value, serving as the objective function. Additionally, model accuracy was assessed using Percent Bias (PBIAS) and Kling-Gupta Efficiency (KGE). The framework outputs include augmented time series data for all variables across the climate, hydrology, and water quality domains, with a monthly frequency. The comprehensive results, including the winning model, the dataset used (original or variant), the corresponding goodness-of-fit indicators, and the respective rating, are presented in Table 4.

Table 4: Imputation results with the winning model, the dataset used (original or variant), the corresponding goodness-of-fit indicators (NSE, KGE, PBIAS), and the respective rating: Very good (VG), Good (G), Satisfactory (S) and Unsatisfactory (U).

Variable	Station	Best model + dataset variant	NSE	NSE ranking	KGE	KGE ranking	PBIAS	PBIAS ranking
P	25 de Agosto	Hubber Regressor (MICE) + EWMA	0.80	G	0.77	G	9.17	G
	San Gabriel	Hubber Regressor (MICE)	0.80	VG	0.78	G	5.71	G
	Reboledo	Hubber Regressor (MICE)	0.83	VG	0.80	G	-1.56	VG
	Cerro Colorado	Ridge + EMA	0.76	G	0.76	G	-1.68	VG
	La Cruz	Bayesian Ridge + SC	0.80	VG	0.81	G	1.10	VG
	Sarandí Grande	Ridge (MICE)	0.73	G	0.73	G	-3.37	VG
	Villa 25 de Mayo	Hubber Regressor + SC	0.89	VG	0.85	G	1.76	VG
	Villa Cardal	KNN + SC	0.86	VG	0.85	G	-0.55	VG
	Mendoza	SVR	0.89	VG	0.89	G	-1.63	VG
h	Florida	KNN + SC	0.98	VG	0.97	G	0.02	VG
Q	Florida	Random Forest Regressor	0.98	VG	0.90	G	-1.85	VG
WT	SLC01	IDW	0.93	VG	0.90	G	-3.14	VG
	SLC02	Hubber Regressor + SC + SD	0.96	VG	0.93	G	0.39	VG
	PS01=SLC03	IDW + SD	0.95	VG	0.95	G	3.77	VG
	PS03	IDW	0.98	VG	0.97	G	-1.21	VG
	PS04	IDW + SD	0.98	VG	0.97	G	1.40	VG
	PS02	IDW	0.97	VG	0.96	G	0.01	VG
DO	SLC01	Hubber Regressor + SC	0.74	G	0.80	G	-0.22	VG
	SLC02	Ridge (MICE) + EWMA	0.81	VG	0.85	G	-0.12	VG
	PS01=SLC03	IDW + SD	0.47	S	0.56	G	-2.31	VG
	PS03	Ridge + EWMA + SC	0.75	G	0.78	G	0.07	VG
	PS04	Hubber Regressor	0.87	VG	0.85	G	-0.01	VG
	PS02	IDW	0.65	S	0.77	G	-1.75	VG
Cond	SLC01	IDW	0.63	S	0.77	G	-1.22	VG
	SLC02	SVR	0.69	S	0.69	G	0.29	VG
	PS01=SLC03	Ridge + SD	0.82	VG	0.86	G	0.77	VG
	PS03	Ridge + EWMA + SC	0.85	VG	0.87	G	1.76	VG
	PS04	IDW	0.97	VG	0.95	G	-1.97	VG
	PS02	Hubber Regressor + SC	0.92	VG	0.91	G	0.04	VG
pH	SLC01	Hubber Regressor + SC	0.45	S	0.52	G	-0.13	VG
	SLC02	SVR (MICE)	0.76	G	0.77	G	-0.03	VG
	PS01=SLC03	SVR + SD	0.49	S	0.52	G	0.09	VG
	PS03	IDW	0.57	S	0.72	G	0.18	VG
	PS04	Ridge + SC	0.79	G	0.80	G	0.00	VG
	PS02	IDW	0.81	VG	0.84	G	-0.41	VG
Turb	SLC01	IDW	0.16	U	0.27	G	12.52	G
	SLC02	Ridge (MICE)	0.53	S	0.63	G	-1.58	VG
	PS01=SLC03	IDW + SD	0.58	S	0.58	G	11.33	G
	PS03	IDW	0.61	S	0.74	G	1.06	VG
	PS04	IDW	0.87	VG	0.90	G	4.27	VG
	PS02	Ridge + SC	0.89	VG	0.89	G	-0.58	VG
BOD	SLC01	TheilSen Regressor + SC	0.38	U	0.39	G	5.55	VG
	SLC02	IDW + SD	0.38	U	0.48	G	-4.61	VG
	PS01=SLC03	SVR	0.21	U	0.18	G	-0.36	VG
NH4+	SLC01	SVR (MICE) + EWMA	0.03	U	0.03	G	11.37	VG
	SLC02	KNN	0.23	U	0.43	G	4.43	VG
	PS01=SLC03	SVR + SD	0.05	U	0.05	G	9.19	VG
	PS03	Hubber Regressor + EWMA + SC	0.80	VG	0.77	G	3.03	VG
	PS04	KNN + SC + SD	0.84	VG	0.80	G	3.69	VG
	PS02	IDW	0.48	S	0.41	G	19.61	G
$NO3^-$	SLC01	Hubber Regressor + SC	0.24	U	0.28	G	22.08	S

Table 4 continued from previous page

Variable	Station	Best model + dataset variant	NSE	NSE ranking	KGE	KGE ranking	PBIAS	PBIAS ranking
	SLC02	Random Forest Regressor + EWMA + SD	0.38	S	0.45	G	-7.69	VG
	PS01=SLC03	Random Forest Regressor + SC	-0.08	U	-0.10	G	-1.22	VG
	PS03	Hubber Regressor + SC	0.60	G	0.69	G	-1.99	VG
	PS04	IDW	0.80	G	0.78	G	-4.14	VG
	PS02	IDW	0.54	G	0.77	G	4.72	VG
NO2-	SLC01	KNN (MICE)	0.50	S	0.51	G	14.67	VG
	SLC02	SVR	0.69	G	0.58	G	-10.89	VG
	PS01=SLC03	SVR	0.21	U	0.26	G	10.70	VG
	PS03	Hubber Regressor + SC	0.75	G	0.74	G	-4.89	VG
	PS04	IDW	0.93	VG	0.89	G	-1.60	VG
	PS02	IDW	0.85	VG	0.83	G	-4.02	VG
TN	SLC01	Random Forest Regressor + SC	0.23	U	0.30	G	1.05	VG
	SLC02	Bayesian Ridge (MICE) + EWMA	0.69	G	0.71	G	-0.32	VG
	PS01=SLC03	Ridge + SC + SD	0.14	U	0.22	G	0.41	VG
	PS03	IDW	0.83	VG	0.90	G	3.31	VG
	PS04	IDW	0.93	VG	0.94	G	-2.49	VG
	PS02	IDW	0.86	VG	0.86	G	1.52	VG
TP	SLC01	Hubber Regressor + SC	-0.04	U	0.02	G	13.22	VG
	SLC02	Random Forest Regressor (MICE) + EWMA	0.16	U	0.29	G	-1.67	VG
	PS01=SLC03	IDW	0.61	G	0.63	G	-3.92	VG
	PS03	IDW	0.80	G	0.80	G	-1.27	VG
	PS04	Hubber Regressor + SC	0.78	G	0.80	G	-0.90	VG
	PS02	Bayesian Ridge (MICE) + EWMA	0.79	G	0.78	G	-0.95	VG
Chl-a	PS01=SLC03	IDW	-0.20	U	-0.21	G	49.34	U
	PS03	KNN + SC	-0.43	U	-0.18	G	1.95	VG
	PS04	Hubber Regressor + EWMA + SC + SD	-0.08	U	0.08	G	23.60	S
	PS02	SVR + EWMA	-0.15	U	-0.37	G	14.61	VG

Considering the NSE rating, the imputation performance for the climatic and hydrometric variables was good overall (NSE>0.73). Regarding the physical water quality variables, on average, adequate imputation performances were obtained. *WT* was the best-imputed variable at the six monitoring stations, reporting very good performance (NSE>0.90). The high daily and annual seasonality that characterizes this variable makes its simulation and, therefore, its imputation less challenging (Rodríguez et al., 2021). The correlations between *WT*-*Cond* and *WT*-*DO* (Figure 5) are reflected in the good performance of such variables at the six monitoring sites. The imputation process for the other water-quality variables (physical and chemical) returned different results depending on the station considered. It is significant to remark that the performance is always outstanding at the three monitoring stations located in the Paso Severino reservoir (PS03, PS04, and PS02), while the imputation can sometimes be unsatisfactory at the sites located upstream of the lake along Santa Lucía Chico river (SLC01, SLC02, and PS01). SLC01 and SLC02 are located several kilometers upstream of the reservoir, where the water body has a fluvial behavior associated with a lotic ecosystem.

While PS02, PS03, and PS04 are located within the lake, where the water body is lacustrine, associated with a lentic ecosystem.

This finding may be due to the different hydrologic response times considering the location of the measurement sites. The hydrograph-base time observed at Florida hydrometric station is overall equal to 6 days, and it generally does not change with the variation of the streamflow magnitude (Rodríguez et al., 2021). Ríos (2019) reported that the Paso Severino renewal time ranges between 2 and 8 weeks. Furthermore, during precipitation events, such renewal time could be a few days long, while it can last several months during dry periods. *Chl-a* and *BOD* were the only two variables that the framework could not adequately impute at any site. They are among the water-quality variables, with data recorded only in three (*BOD*) or four stations (*Chl-a*). This means the spatial constraints related to the variant SD and SC are limited. Furthermore, the correlated variables resulted from the three correlation matrices (Pearson, Spearman, and Kendall) and from the variable-dependency tree (Figure 5) that were supposed to aid the imputation of *Chl-a* and *BOD* were very few. This supports the reliability of the domain constraints implemented in the framework.

The validation of the imputation process was notable, showing overall very good and good results in terms of the PBIAS and KGE ratings. A box-plot representation of the framework NSE performance per domain is represented in Figure 7. Additionally, box plots of the framework PBIAS and KGE performance are reported in the Supplementary Information (A.3).

More than 75% of the imputed data is characterized by NSE>0.45 (satisfactory results). In particular, the minimum NSE computed for meteorologic variables is 0.72, meaning that all the imputations can be considered good (33%) and very good (66%). For hydrometric variables, NSE is always >0.97, showing the very good performance of the proposed framework. The performance tends to decrease when dealing with water-quality variables. More than 78% of the physical-water-quality variables are characterized by NSE>0.45 (satisfactory results), and more than 66% of the chemical-water quality variables reach NSE>0.35 (satisfactory results). For both domains, more than 91% of the imputed data has NSE>0, meaning that for almost all the water-quality imputations, the proposed framework is better than the mean function used as an imputer.

4.2. Selection of the best dataset-model pair

In this study, we considered various model and dataset options in our framework beyond the original ones. Table 5 displays how often each model with corresponding variants was chosen and Table 6 summarizes the

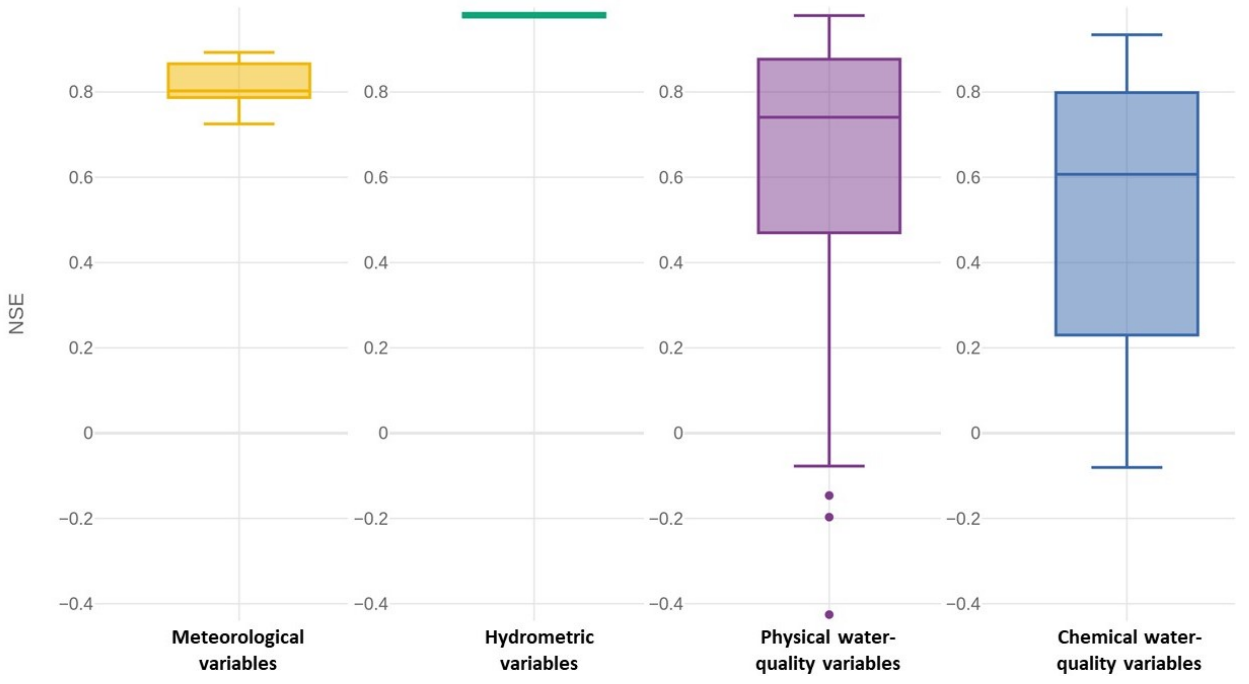


Figure 7: Box plots depicting the framework's performance in terms of Nash-Sutcliffe Efficiency (NSE) are shown for variables within the meteorological, hydrometric, physical-water quality, and chemical-water quality domains.

number of times a dataset and model variant was selected to impute a variable. It is crucial to note that there is no superior model for a specific variable type or domain (Wolpert and Macready, 1997). This emphasizes the significance of a framework like ours, where multiple models are implemented and run to achieve optimal imputation performance in varying scenarios.

From Table 5, it is clear that IDW is a successful imputation technique, particularly for water-quality variables (chosen 27 times). It returns outstanding results for the variables recorded at the three stations located in the Paso Severino reservoir (especially for *WT*, *Turb*, and *TN*), and it is sometimes chosen as the best model for *WT*, *DO*, *Cond*, *Turb*, *BOD*, *TP*, and *Chl-a* at the stations located along SLC river. It is interesting to see that IDW was always chosen in its original form (22 times) or with the SD variant (5 times) (Table 5). The HR, RR, and TR models are similar linear methods that differ from each other only in their training techniques. They were also chosen as best models 27 times (29.3% of the time) as well as IDW.

The MICE-model variant was selected 12 times as the best model. This confirmed the results by Jones et al. (2014), where they report that MICE can effectively fill in missing values in water-quality data, and the findings reported by Ratolojanahary et al. (2019), where they state that the hybridization of MICE with

Table 5
Number of times each model was selected as the best model to impute a variable.

Imputation model	Dataset and/or model variant	Number of imputed variables	Total
Inverse Distance Weighting (IDW)	Original	22	27
	SD	5	
Hubber Regressor (HR)	SC	9	16
	MICE	2	
	SC + SD	1	
	EWMA + SC	1	
	Original	1	
	MICE + EWMA	1	
	EWMA + SC + SD	1	
Ridge Regressor (RR)	EWMA + SC	2	10
	MICE	2	
	SC	2	
	MICE + EWMA	1	
	EWMA	1	
	SC + SD	1	
	SD	1	
Support Vector Regressor (SVR)	Original	5	10
	SD	2	
	MICE + EWMA	1	
	MICE	1	
	EWMA	1	
K-nearest Neighbors Regressor (KNNR)	SC	3	6
	Original	1	
	SC + SD	1	
	MICE	1	
Random Forest Regressor (RFR)	SC	2	5
	Original	1	
	EWMA + SD	1	
	MICE + EWMA	1	
Bayesian Ridge Regressor (BRR)	MICE + EWMA	2	3
	SC	1	
TheilSen Regressor (TR)	SC	1	1

several machine-learning algorithms (SVR, KNN, RF, and boosted regression tree) always performs better than the original MICE taken alone. It is worth noting that, in our work, the MICE variant was always selected either with the original dataset or with the EWMA-dataset variant (Table 5).

The data presented in Table 6 clearly shows that the best model selection occurred a whopping 62 times, which represents over 67% of the total selection instances. This undoubtedly emphasizes the crucial role of incorporating physical constraints in the machine-learning framework to enhance imputation performance. Notably, dataset variants related to spatial variability were preferred over those related to temporal variability,

Table 6
Number of times each model variant was selected to impute a variable.

Variant		N° of imputed variables
Dataset and model	Original	30
Dataset	SC	25
	SD	13
	EWMA	12
Model	MICE	12

with remarkable selection rates of 41.3% and 13%, respectively. SC was the most frequently selected among the spatial variants, accounting for 27.2% of the total selection instances. These results undoubtedly call for further in-depth exploration of the subject.

The integration of interpretable machine learning models within the proposed data imputation framework presents a forward-looking perspective, enhancing the transparency and understanding of the imputation process for environmental data encompassing meteorology, hydrology, and water quality. This shift towards interpretability fosters trust and credibility by demystifying the model's decision-making, allowing for insightful evaluation of the framework's efficacy. Researchers benefit from the identification of environmental patterns and actionable insights, while stakeholders receive clear explanations for imputed values, crucial for informed decision-making. The interpretable nature of the framework aligns with scientific principles, ensuring adherence to known physical processes and reinforcing its applicability in real-world contexts. Overall, the emphasis on interpretability adds depth and transparency to the proposed framework, positioning it as a robust and insightful approach to environmental data imputation.

4.3. On the value of the proposed framework

The proposed framework exhibits three distinctive features that contribute to its effectiveness. Firstly, it adopts a multi-domain approach, which sets it apart from previous studies. By simultaneously imputing environmental variables from different domains, it becomes a powerful tool for enhancing the performance of integrated complex models at the catchment scale through data imputation. This unique characteristic enables comprehensive analysis and improves the accuracy of predictions.

Secondly, the framework incorporates physical constraints, combining machine learning with domain-specific knowledge. This incorporation ensures that imputations align with the known physical processes, enhancing the reliability and interpretability of results. The latter demonstrates the advantages of this

approach compared to purely data-driven techniques. A notable comparison can be made with our previous work (Rodríguez et al., 2021) where we evaluated the performance of different machine learning algorithms for water-quality imputation. The physically-constrained framework consistently outperformed the pure data-driven models, highlighting the value of integrating physical constraints.

It is important to note that the physical constraints utilized in this study, such as variable correlation and spatial variability, were specifically designed for our study site. However, they can be adapted and applied to other geographical regions, showcasing the framework's generalization capability. This flexibility allows the framework to be effectively utilized in diverse watershed scenarios.

Thirdly, the proposed framework addresses the challenge of model selection, acknowledging the inherent uncertainty in choosing a single machine-learning model. Unlike a simplistic approach of running a single model, the framework systematically evaluates and compares the performance of multiple machine learning and statistical models. By assessing various algorithms and configurations, it aims to identify the most suitable model for imputing environmental missing data under diverse conditions.

Furthermore, the proposed approach consolidates the positive aspects observed in previous studies. It effectively handles a high percentage of missing values and incorporates a wide range of statistical and machine-learning techniques, as observed in various works (Aguilera et al., 2020; Chen et al., 2021; Jones et al., 2014; Ratolojanahary et al., 2019,?; Zhang and Thorburn, 2022). The framework offers a comprehensive and versatile solution for data imputation tasks by encompassing these advantageous elements.

Our framework significantly contributes to advancing the understanding of the environmental system, addressing both direct and indirect aspects. Through direct contributions, it enhances model accuracy and parameter optimization and facilitates improved predictive modeling and hypothesis testing in environmental science. Policymakers benefit directly by gaining access to accurate and complete environmental data, enabling the development of precise, data-driven policies for real-world challenges. Indirectly, the framework minimizes biases from incomplete datasets, fostering a robust foundation for environmental studies and enriching our overall understanding. The imputed data supports a holistic view of environmental variables, contributing to a broader knowledge base. Additionally, the framework's impact extends over time, accumulating reliable data to strengthen the scientific foundation of long-term environmental research.

5. Conclusions

In this study, we have developed a novel framework that effectively addresses the challenge of imputing missing data in various environmental domains, including meteorology, hydrology, and water quality. This framework combines data-driven models with physical knowledge, resulting in satisfactory imputation results. The key features of this framework are as follows:

i) It incorporates physical constraints such as variable correlations (Pearson, Spearman, and Kendall correlation matrices), temporal variability (EWMA), and spatial variability (SD and SC) of the features under study. By considering these constraints, the framework ensures that the imputed data aligns with the underlying physical characteristics of the variables.

ii) The framework demonstrates a high success rate in imputing a substantial percentage of missing data, surpassing 70%. This ability to handle a large proportion of missing values enhances the overall data completeness.

iii) It adopts a multivariate approach, simultaneously considering various variables. This multivariate aspect allows for comprehensive analysis and improves the accuracy of the imputed data.

iv) The framework incorporates diverse statistical and machine-learning methods, contributing to flexibility and robustness. The framework can effectively capture the complex relationships within the data by employing various techniques.

The framework's performance was rigorously evaluated through cross-validation, selecting the best model for each variable. Overall, the results were satisfactory, with minimum Nash-Sutcliffe Efficiency (NSE) values above 0.72 for meteorologic variables, indicating good to very good imputations. Hydrometric variables consistently achieved NSE values above 0.97, demonstrating excellent performance. Water-quality variables exhibited slightly lower NSE values, but over 78% of the physical-water-quality variables and 66% of the chemical-water quality variables reached satisfactory NSE levels.

Regarding model selection, the Inverse Distance Weighting (IDW) method was particularly effective for imputing water-quality variables. In contrast, linear methods such as Historical Records (HR), Regression Relations (RR), and Transfer Relations (TR) were also successful. The study highlights that no single best model per variable type or domain exists, underscoring the importance of employing a framework rather than relying on individual models. Furthermore, more than 67% of the time, a variant or combination of variants was identified as the best-selected model, emphasizing the significance of incorporating physical

knowledge into the framework. In our case study, dataset variants related to spatial variability were selected more frequently than those related to temporal variability (41.3% and 13%, respectively).

The outcomes of this study are expected to contribute significantly to the accurate imputation and augmentation of environmental data. Integrating such data into watershed-scale models will enhance the performance of water-quality simulations and predictions, enabling improved decision-making in various applications.

To further improve the framework, it is crucial to highlight its limitations, which will be the gaps where to focus future research. Upstream of the reservoir, changes in hydrological conditions, such as flow rates or pollutant sources, may be more dynamic and less predictable. The physics-based constraints in the framework might not fully account for the complexities of upstream hydrological and water quality processes. This incomplete understanding could lead to inaccuracies in imputing water quality missing values in these regions. Another weakness of the framework is represented by the availability and quality of auxiliary data used for training it. Its effectiveness depends on them. In situations where relevant auxiliary data are scarce or unreliable, the imputation accuracy may be compromised. Understanding these limitations and tailoring the framework to address these specific challenges is essential for improving its overall performance and ensuring accurate imputations in diverse environmental settings.

6. Software and data availability

The source datasets used in this work are available for reuse [dataset] Environmental data imputation project (2022a). They are published as four PARQUET files: i) CA_DINACEA_2004_2020 (water quality variables), ii) HIDRO_DINAGUA_1971_2020 (hydrometric variables), iii) MET_INIA_2013_2020 and iv) MET_INUMET_1980_2020 (meteorological variables) with a total size of 1.11 MB.

The datasets obtained after applying the imputation methodology described in this work are also available [dataset] Environmental data imputation project (2022b). We provide four PARQUET files corresponding to the original datasets, which are available with a total size of 0.21 MB.

The data imputation framework developed for this work is freely available at `https://gitlab.com/fing-hydroinformatics/fsda-lu-wq`. Models devised in this work can be accessed from `https://gitlab.com/fing-hydroinformatics/fsda-lu-wq/-/tree/paper`. The framework is implemented using *Python* 3.10 and can be executed using *docker-compose* on any general-purpose computer.

Funding

This work was partially supported by the National Research and Innovation Agency (ANII) [grant number FSDA-1-2018-1-153967].

CRediT authorship contribution statement

Marcos Pastorini: Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Review & Editing, Visualization. **Rafael Rodríguez:** Resources, Methodology, Formal analysis, Investigation, Writing - Review & Editing. **Lorena Etcheverry:** Conceptualization, Supervision, Writing - Review & Editing. **Alberto Castro:** Conceptualization, Investigation, Supervision, Writing - Review & Editing, Funding acquisition. **Angela Gorgoglione:** Conceptualization, Investigation, Resources, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration, Funding acquisition.

A. Supplementary information

A.1. Complete dataset

A summary of the complete dataset used in this study is presented in Table 7 (meteorological and hydrometric variables) and Table 8 (water quality variables), where the percentage of missing values detected for each variable (% N/A) is also reported.

Table 8: Summary of the water quality dataset.

Water Quality Dataset. Source: DINACEA

Variable	Station	Frequency	Period	% N/A
WT	SLC01	Monthly	30/9/2014 - 31/8/2020	51.4
	SLC02		30/9/2014 - 31/8/2020	51.4
	PS01=SLC03		30/9/2014 - 31/8/2020	63.9
	PS02		30/9/2014 - 31/8/2020	61.1
	PS03		30/9/2014 - 31/8/2020	59.7
	PS04		30/9/2014 - 31/8/2020	59.7
Cond	SLC01	Monthly	30/9/2014 - 31/8/2020	51.4
	SLC02		30/9/2014 - 31/8/2020	51.4
	PS01=SLC03		30/9/2014 - 31/8/2020	63.9
	PS02		30/9/2014 - 31/8/2020	59.7

Continued on next page

Table 8 – continued from previous page

Water Quality Dataset. Source: DINACEA

Variable	Station	Frequency	Period	% N/A
	PS03		30/9/2014 - 31/8/2020	59.7
	PS04		30/9/2014 - 31/8/2020	59.7
NO2-, NO3-, NH4+	SLC01	Monthly	30/9/2014 - 31/8/2020	51.4
	SLC02		30/9/2014 - 31/8/2020	51.4
	PS01=SLC03		30/9/2014 - 31/8/2020	63.9
	PS02		30/9/2014 - 31/8/2020	59.7
	PS03		30/9/2014 - 31/8/2020	61.1
	PS04		30/9/2014 - 31/8/2020	61.1
TN	SLC01	Monthly	30/9/2014 - 31/8/2020	52.8
	SLC02		30/9/2014 - 31/8/2020	52.8
	PS01=SLC03		30/9/2014 - 31/8/2020	65.3
	PS02		30/9/2014 - 31/8/2020	61.1
	PS03		30/9/2014 - 31/8/2020	62.5
	PS04		30/9/2014 - 31/8/2020	62.5
DO	SLC01	Monthly	30/9/2014 - 31/8/2020	51.4
	SLC02		30/9/2014 - 31/8/2020	51.4
	PS01=SLC03		30/9/2014 - 31/8/2020	63.9
	PS02		30/9/2014 - 31/8/2020	59.7
	PS03		30/9/2014 - 31/8/2020	59.7
	PS04		30/9/2014 - 31/8/2020	59.7
Turb	SLC01	Monthly	30/9/2014 - 31/8/2020	52.8
	SLC02		30/9/2014 - 31/8/2020	52.8
	PS01=SLC03		30/9/2014 - 31/8/2020	65.3
	PS02		30/9/2014 - 31/8/2020	61.1
	PS03		30/9/2014 - 31/8/2020	62.5
	PS04		30/9/2014 - 31/8/2020	62.5
PO43-	SLC01	Monthly	31/10/2018 - 31/8/2020	47.8
	SLC02		31/10/2018 - 31/8/2020	47.8
	PS01=SLC03		31/10/2018 - 31/8/2020	47.8
	PS02		31/10/2018 - 31/8/2020	73.9
	PS03		31/10/2018 - 31/8/2020	69.6

Continued on next page

Table 8 – continued from previous page

Water Quality Dataset. Source: DINACEA

Variable	Station	Frequency	Period	% N/A
	PS04		31/10/2018 - 31/8/2020	69.6
TP	SLC01	Monthly	31/10/2018 - 31/8/2020	51.4
	SLC02		31/10/2018 - 31/8/2020	51.4
	PS01=SLC03		31/10/2018 - 31/8/2020	65.3
	PS02		31/10/2018 - 31/8/2020	62.5
	PS03		31/10/2018 - 31/8/2020	61.1
	PS04		31/10/2018 - 31/8/2020	61.1
Chl-a	PS01=SLC03	Monthly	31/10/2018 - 31/8/2020	63.9
	PS02		31/10/2018 - 31/8/2020	61.1
	PS03		31/10/2018 - 31/8/2020	62.5
	PS04		31/10/2018 - 31/8/2020	62.5
BOD	SLC01	Monthly	31/10/2018 - 31/8/2020	51.4
	SLC02		31/10/2018 - 31/8/2020	51.4
	PS01=SLC03		31/10/2018 - 31/8/2020	56.9
	PS02		31/10/2018 - 31/8/2020	71.4
	PS03		31/10/2018 - 31/8/2020	71.4
	PS04		31/10/2018 - 31/8/2020	71.4
pH	SLC01	Monthly	31/10/2018 - 31/8/2020	52.8
	SLC02		31/10/2018 - 31/8/2020	52.8
	PS01=SLC03		31/10/2018 - 31/8/2020	65.3
	PS02		31/10/2018 - 31/8/2020	61.1
	PS03		31/10/2018 - 31/8/2020	61.1
	PS04		31/10/2018 - 31/8/2020	61.1
TS	SLC01	Monthly	31/10/2018 - 31/8/2020	47.8
	SLC02		31/10/2018 - 31/8/2020	47.8
	PS01=SLC03		31/10/2018 - 31/8/2020	47.8
	PS02		31/10/2018 - 31/8/2020	69.6
	PS03		31/10/2018 - 31/8/2020	69.6
	PS04		31/10/2018 - 31/8/2020	69.6
TSS	SLC01	Monthly	31/10/2018 - 31/8/2020	47.8
	SLC02		31/10/2018 - 31/8/2020	47.8
	PS01=SLC03		31/10/2018 - 31/8/2020	47.8

Continued on next page

Table 8 – continued from previous page

Water Quality Dataset. Source: DINACEA

Variable	Station	Frequency	Period	% N/A
	PS02		28/2/2018 - 31/8/2020	74.2
	PS03		31/10/2018 - 31/8/2020	69.6
	PS04		31/10/2018 - 31/8/2020	69.6

A.2. Variable correlation and results

Table 9 provides the scientific literature that supports each correlation reported in Figure 5.

For the sake of completeness, Pearson and Kendall's matrices are reported in Figure 8 Figure and 9.

The complete list of the variable correlation resulting from the Pearson, Spearman, and Kendall matrices is reported below:

'Precipitation': ['Chlorophyll-a', 'Nitrite', 'Potential of hydrogen', 'Precipitation', 'Total nitrogen'],

'Evapotranspiration': ['Average air temperature', 'Average relative humidity', 'Chlorophyll-a', 'Dissolved oxygen', 'Evapotranspiration', 'Heliophany', 'Maximum air temperature', 'Minimum air temperature', 'Solar Radiation', 'Water temperature'],

'Maximum air temperature': ['Average air temperature', 'Chlorophyll-a', 'Dissolved oxygen', 'Evapotranspiration', 'Heliophany', 'Maximum air temperature', 'Minimum air temperature', 'Solar Radiation', 'Turbidity', 'Water temperature'],

'Average air temperature': ['Average air temperature', 'Chlorophyll-a', 'Conductivity', 'Dissolved oxygen', 'Evapotranspiration', 'Maximum air temperature', 'Minimum air temperature', 'Solar Radiation', 'Turbidity', 'Water temperature'],

'Minimum air temperature': ['Average air temperature', 'Chlorophyll-a', 'Conductivity', 'Dissolved oxygen', 'Evapotranspiration', 'Maximum air temperature', 'Minimum air temperature', 'Turbidity', 'Water temperature'],

'Solar Radiation': ['Average air temperature', 'Average relative humidity', 'Dissolved oxygen', 'Evapotranspiration', 'Heliophany', 'Maximum air temperature ', 'Solar Radiation', 'Water temperature'],

'Heliophany': ['Average relative humidity', 'Dissolved oxygen', 'Evapotranspiration', 'Heliophany', 'Maximum air temperature', 'Solar Radiation', 'Water temperature'],

Table 7
Summary of the meteorological and hydrometric datasets.

Source	Variable	Station	Frequency	Period	% N/A
Metereological Datasets					
INIA	*ET, RH, Tave , Tave , Tmax , Tmin , Hel , SR, WS*	Las Brujas	Once a day	1/8/2014 – 31/12/2020	0
INUMET	*P*	Florida	Once a day	1/8/2014 – 29/6/2020	0
		Reboledo		1/8/2014 – 29/6/2020	23
		San Gabriel		1/8/2014 – 29/6/2020	3.2
		Villa 25 de Mayo		2/8/2014 – 20/2/2019	3.3
		Mendoza		1/8/2014 – 29/6/2020	1.8
		Cerro Colorado		1/8/2014 – 29/6/2020	1.6
		Sarandí Grande		1/8/2014 – 29/6/2020	1.3
		La Cruz		1/8/2014 – 29/6/2020	1.9
		Villa Cardal		1/8/2014 – 29/6/2020	1.1
		25 de Agosto		1/8/2014 – 29/6/2020	1.7
Hydrometric dataset					
DINAGUA	*Q , h*	Florida	Three times a day	1/8/2014 – 30/6/2020	5.6

'Average relative humidity': ['Average relative humidity', 'Dissolved oxygen', 'Evapotranspiration', 'Heliophany', 'Solar Radiation'],

'Wind speed': ['Wind speed'],

'Streamflow': ['Chlorophyll-a', 'Nitrate', 'Streamflow', 'Water level'],

'Water level': ['Chlorophyll-a', 'Streamflow', 'Turbidity', 'Water level'],

'Water temperature': ['Average air temperature', 'Chlorophyll-a', 'Conductivity', 'Dissolved oxygen', 'Evapotranspiration', 'Heliophany', 'Maximum air temperature', 'Minimum air temperature', 'Solar Radiation', 'Total phosphorus', 'Turbidity', 'Water temperature'],

'Conductivity': ['Average air temperature', 'Chlorophyll-a', 'Conductivity', 'Dissolved oxygen', 'Glyphosate', 'Minimum air temperature', 'Nitrate', 'Nitrite', 'Potential of hydrogen', 'Total nitrogen', 'Total phosphorus', 'Turbidity', 'Water temperature'],

'Dissolved oxygen': ['Average air temperature ', 'Average relative humidity', 'Chlorophyll-a', 'Conductivity', 'Dissolved oxygen', 'Evapotranspiration', 'Heliophany', 'Maximum air temperature', 'Minimum air temperature', 'Solar Radiation', 'Total nitrogen', 'Total phosphorus', 'Turbidity', 'Water temperature'],

'Potential of hydrogen': ['Biochemical oxygen demand', 'Conductivity', 'Potential of hydrogen', 'Precipitation', 'Total nitrogen', 'Total phosphorus', 'Turbidity'],

Table 9
Scientific literature supporting correlations presented in Figure 5.

Correlation	Reference
Solar radiation - Water temperature	Shinohara et al. (2021)
Heliophany - Water temperature	Shinohara et al. (2021)
Air temperature - Water temperature	Shinohara et al. (2021)
Solar radiation - Clorophyll-a	Poll et al. (2021)
Heliophany - Clorophyll-a	Villate et al. (2008)
Air temperature - Clorophyll-a	Villate et al. (2008)
Air temperature - Turbidity	Gorgoglione et al. (2020b)
Precipitation - Water level	Chen et al. (2020)
Precipitation - Turbidity	Gorgoglione et al. (2020b)
Evapotranspiration - Water level	Zhang and Wang (2021)
Evapotranspiration - Turbidity	Gorgoglione et al. (2020b)
Water temperature - Dissolved oxygen	Gorgoglione et al. (2020b); Rodríguez et al. (2021)
Water temperature - Conductivity	Hayashi (2004); Paaijmans et al. (2008)
Water temperature - Clorophyll-a	Crisci et al. (2017); Haakonsson et al. (2017)
Water level - Streamflow	Ye et al. (2017)
Streamflow - Ammonium	Gorgoglione et al. (2020b)
Streamflow - Nitrite	Gorgoglione et al. (2020b)
Streamflow - Nitrate	Gorgoglione et al. (2020b)
Streamflow - Turbidity	Göransson et al. (2013)
Streamflow - Total Nitrogen	Song et al. (2022)
Streamflow - Clorophyll-a	Acker (2005)
Streamflow - Total Phosphorus	Ellison and Brett (2006)
Ammonium - Clorophyll-a	Iriarte et al. (2007)
Ammonium - Total Nitrogen	Iriarte et al. (2007); Satpathy et al. (2011)
Nitrite - Clorophyll-a	Balachandran et al. (1989)
Nitrite - Total Nitrogen	Allott et al. (1995)
Nitrate - Clorophyll-a	Gong et al. (2000)
Nitrate - Total Nitrogen	Allott et al. (1995)
Turbidity - Conductivity	Bakhtiar Jemily et al. (2019)
Turbidity - Total Nitrogen	Lintern et al. (2018)
Turbidity - Total Phosphorus	Villa et al. (2019)
Turbidity - Clorophyll-a	Crisci et al. (2017)
Conductivity - pH	Saalidong et al. (2022)
Total Nitrogen - Clorophyll-a	Bennett et al. (2021); Kärcher et al. (2020)
Total Phosphorus - Clorophyll-a	Bennett et al. (2021); Kärcher et al. (2020)

'Turbidity': ['Average air temperature', 'Chlorophyll-a', 'Conductivity', 'Dissolved oxygen', 'Maximum air temperature', 'Minimum air temperature ', 'Nitrate', 'Nitrite', 'Potential of hydrogen', 'Total nitrogen', 'Total phosphorus', 'Turbidity', 'Water level', 'Water temperature'],

'Biochemical oxygen demand': ['Biochemical oxygen demand', 'Potential of hydrogen'],

'Chlorophyll-a': ['Average air temperature', 'Chlorophyll-a', 'Conductivity', 'Dissolved oxygen', 'Evapotranspiration', 'Maximum air temperature ', 'Minimum air temperature', 'Nitrate', 'Nitrite', 'Precipitation', 'Streamflow', 'Total phosphorus', 'Turbidity', 'Water level', 'Water temperature'],

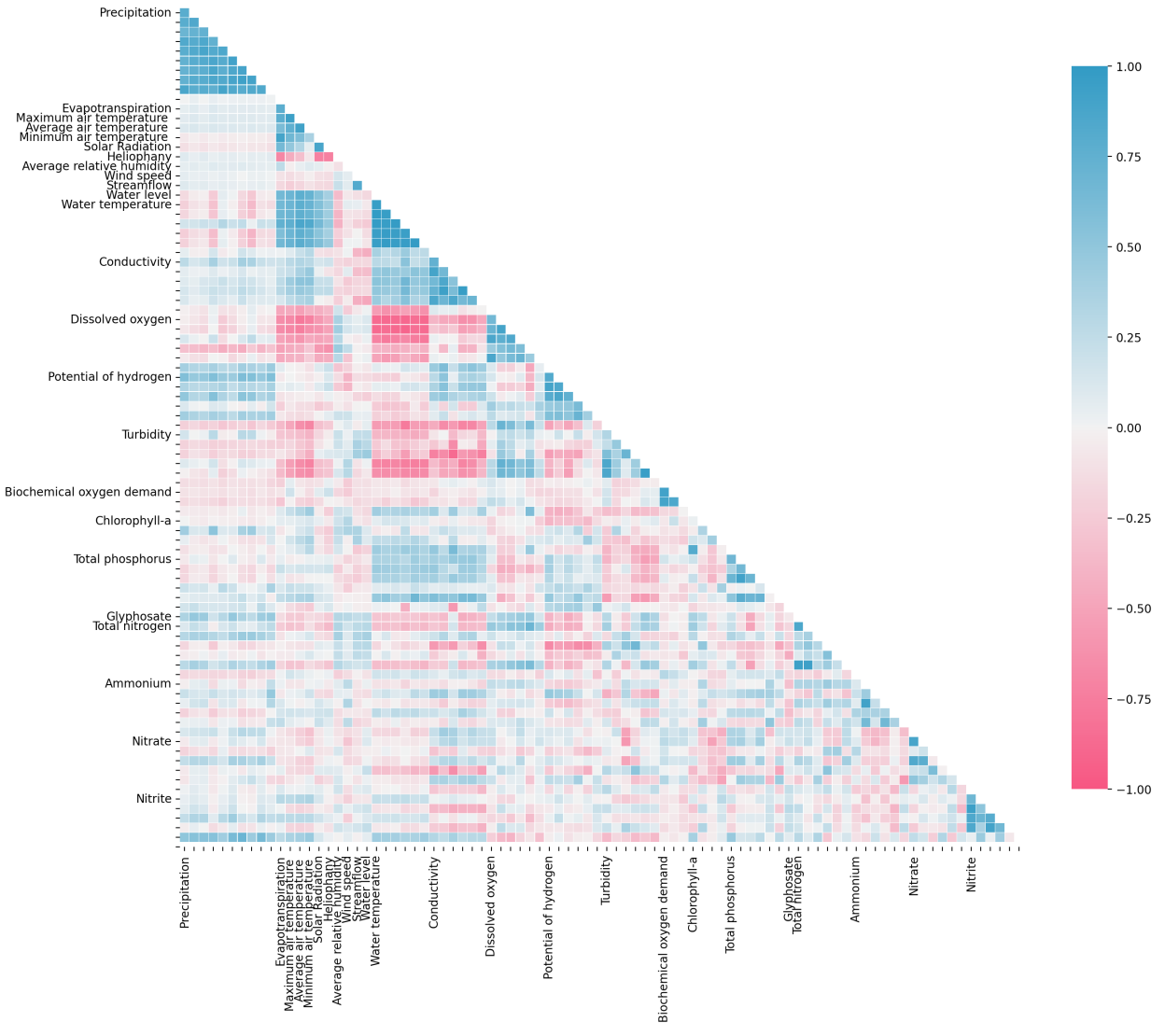


Figure 8: Pearson correlation matrix.

'Total phosphorus': ['Ammonium', 'Chlorophyll-a', 'Conductivity', 'Dissolved oxygen', 'Potential of hydrogen', 'Total nitrogen', 'Total phosphorus', 'Turbidity', 'Water temperature'],

'Glyphosate': ['Conductivity', 'Glyphosate', 'Nitrite'], 'Total nitrogen': ['Conductivity', 'Dissolved oxygen', 'Nitrate', 'Potential of hydrogen', 'Precipitation', 'Total nitrogen', 'Total phosphorus', 'Turbidity'],

'Ammonium': ['Ammonium', 'Nitrate', 'Total phosphorus'],

'Nitrate': ['Ammonium', 'Chlorophyll-a', 'Conductivity', 'Nitrate', 'Streamflow', 'Total nitrogen', 'Turbidity'],

'Nitrite': ['Chlorophyll-a', 'Conductivity', 'Glyphosate', 'Nitrite', 'Precipitation', 'Turbidity'].

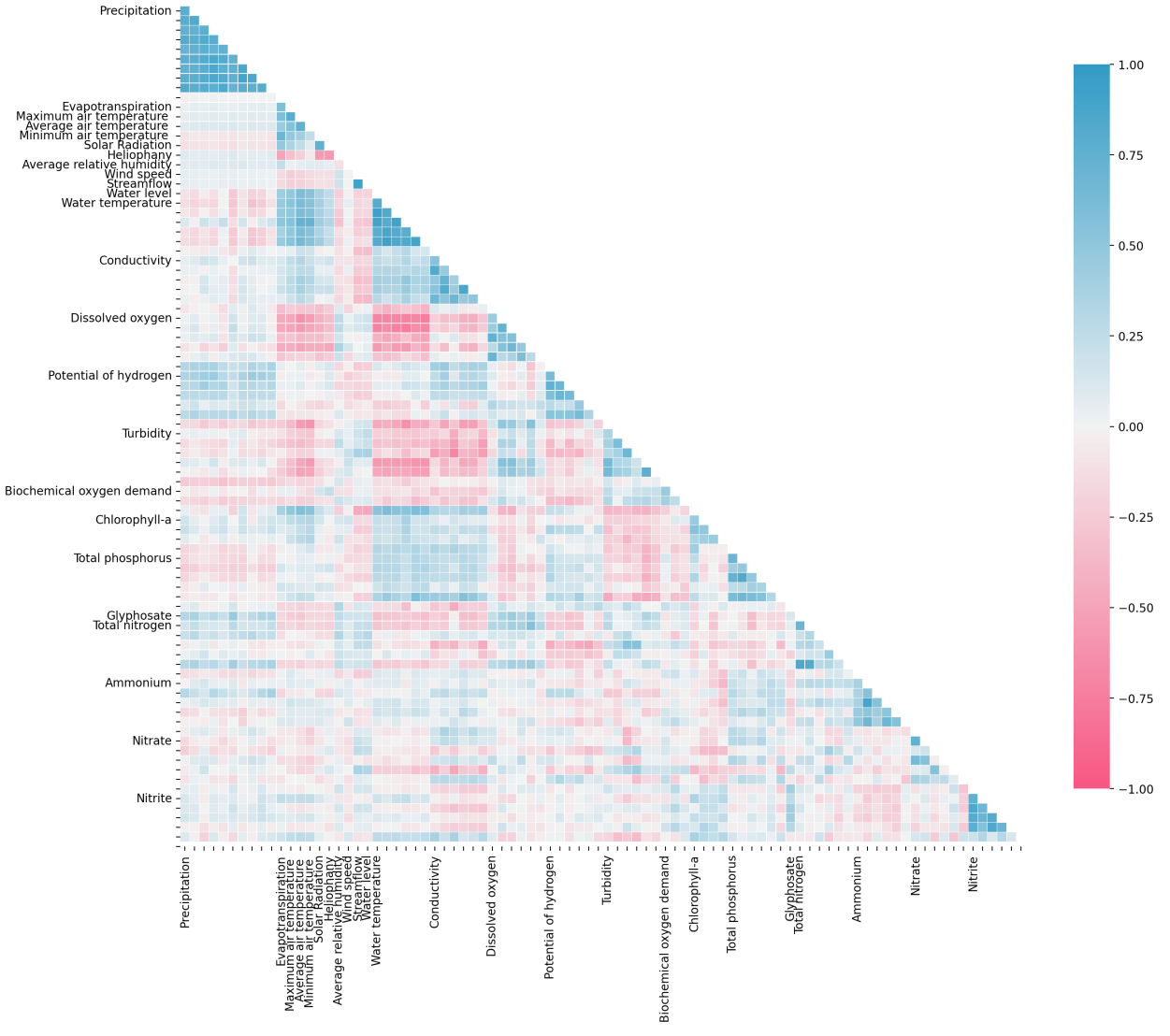


Figure 9: Kendall correlation matrix.

A.3. Complementary results

Box plots of the framework PBIAS and KGE performance are reported in Figures 10 and 11.

References

Acker, J.G., 2005. Remotely-sensed chlaat the Chesapeake Bay mouth is correlated with annual freshwater flow to Chesapeake Bay. Geophysical research letters 32. URL: http://dx.doi.org/10.1029/2004gl021852, doi:{10.1029/2004gl021852}.

Aguilera, H., Guardiola-Albert, C., Serrano-Hidalgo, C., 2020. Estimating extremely large amounts of missing precipitation data. Journal of hydroinformatics 22, 578–592. URL: http://dx.doi.org/10.2166/hydro.2020.127, doi:{10.2166/hydro.2020.127}.

Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M., 2019. Optuna: A next-generation hyperparameter optimization framework, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, ACM, New York, NY, USA.

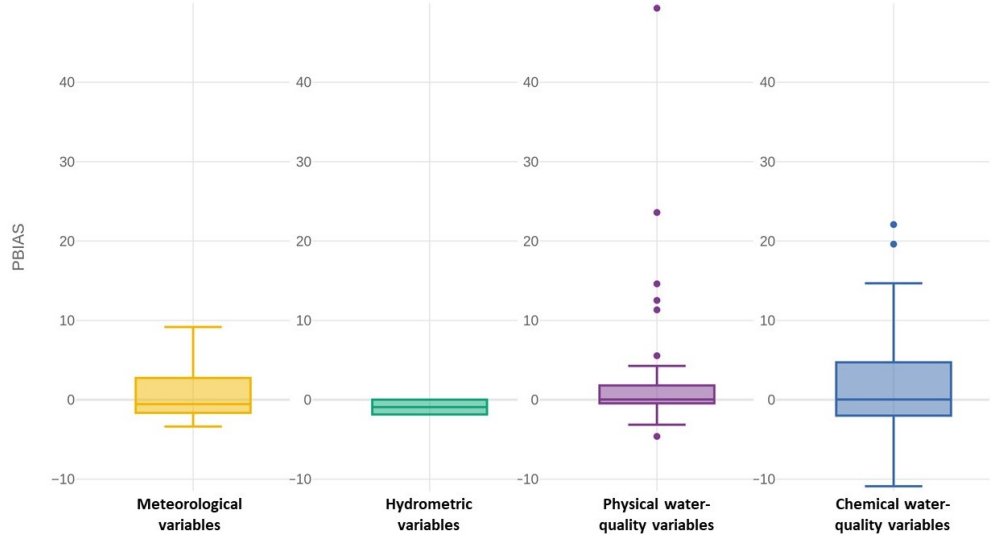


Figure 10: Box plots of the framework PBIAS performance for the variables belonging to the meteorological, hydrometric, physical-water quality, and chemical-water quality domain.

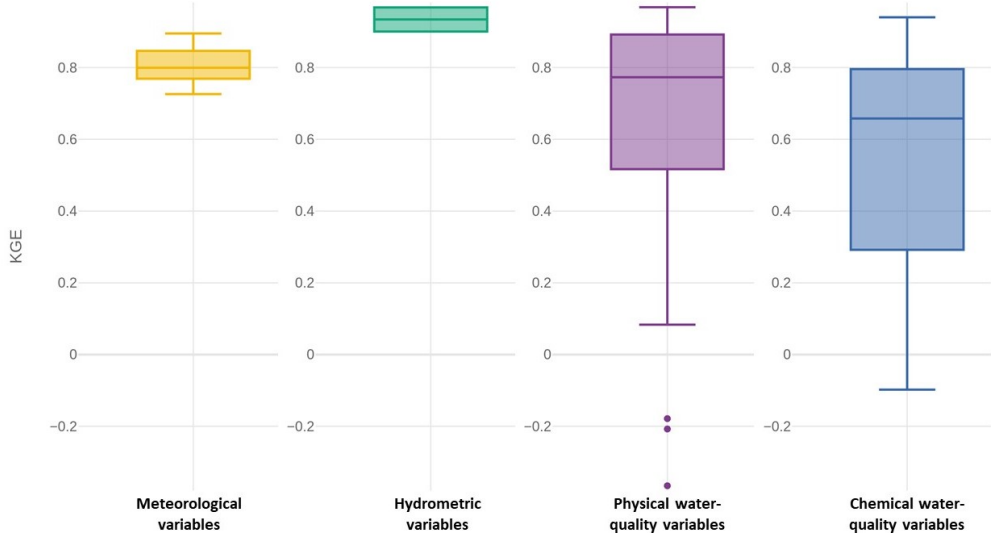


Figure 11: Box plots of the framework KGE performance for the variables belonging to the meteorological, hydrometric, physical-water quality, and chemical-water quality domain.

Allott, T.E.H., Curtis, C.J., Hall, J., Harriman, R., Battarbee, R.W., 1995. The impact of nitrogen deposition on upland surface waters in Great Britain: A regional assessment of nitrate leaching. Water, air, and soil pollution 85, 297–302. URL: http://dx.doi.org/10.1007/bf00476845, doi:{10.1007/bf00476845}.

Andridge, R.R., Little, R.J.A., 2010. A review of hot deck imputation for survey non-response. Revue internationale de statistique [International statistical review] 78, 40–64. URL: http://www.jstor.org/stable/27919794, doi:{10.1111/j.1751-5823.2010.00103.x}.

Azur, M.J., Stuart, E.A., Frangakis, C., Leaf, P.J., 2011. Multiple imputation by chained equations: what is it and how does it work?: Multiple imputation by chained equations. International journal of methods in psychiatric research 20, 40–49. URL: http://dx.doi.org/10.1002/mpr.329, doi:{10.1002/mpr.329}.

Bakhtiar Jemily, N.H., Ahmad Sa'ad, F.N., Mat Amin, A.R., Othman, M.F., Mohd Yusoff, M.Z., 2019. Relationship between electrical conductivity and total dissolved solids as water quality parameter in teluk lipat by using regression analysis. Springer International Publishing, Cham. p. 169–173.

Balachandran, V.K., Rajagopalan, M., Pillai, V.K., 1989. Chlorophyll a and pheo - pigment as indices of biological productivity in the inshore surface waters off Cochin. Indian Journal of Fisheries 36, 227–237. URL: http://eprints.cmfri.org.in/315/.

Bennett, M.G., Lee, S.S., Schofield, K.A., Ridley, C.E., Washington, B.J., Gibbs, D.A., 2021. Response of chlorophyll a to total nitrogen and total phosphorus concentrations in lotic ecosystems: a systematic review. Environmental evidence 10. URL: http://dx.doi.org/10.1186/s13750-021-00238-8, doi:{10.1186/s13750-021-00238-8}.

Bertsimas, D.J., Pawlowski, C., Zhuo, Y.D., 2018. From predictive methods to missing data imputation: An optimization approach. Journal of machine learning research: JMLR , 1–39URL: https://dspace.mit.edu/handle/1721.1/130111?show=full.

Bi, J., Wang, Z., Yuan, H., Ni, K., Qiao, J., 2022. Multi-indicator water time series imputation with autoregressive generative adversarial networks, in: 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 2003–2008. doi:10.1109/SMC53654.2022.9945604.

Blenkinsop, S., Lewis, E., Chan, S.C., Fowler, H.J., 2017. Quality-control of an hourly rainfall dataset and climatology of extremes for the UK: QUALITY-CONTROL AND EXTREMES CLIMATOLOGY FOR UK HOURLY RAINFALL. International journal of climatology: a journal of the Royal Meteorological Society 37, 722–740. URL: http://dx.doi.org/10.1002/joc.4735, doi:{10.1002/joc.4735}.

Breiman, L., 2001. Random Forests. Machine learning 45, 5–32. URL: http://dx.doi.org/10.1023/a:1010933404324, doi:{10.1023/a:1010933404324}.

Bø, T.H., Dysvik, B., Jonassen, I., 2004. LSimpute: accurate estimation of missing values in microarray data with least squares methods. Nucleic acids research 32, e34. URL: http://dx.doi.org/10.1093/nar/gnh026, doi:{10.1093/nar/gnh026}.

Chandra, R., Cripps, S., Butterworth, N., Muller, R.D., 2021. Precipitation reconstruction from climate-sensitive lithologies using Bayesian machine learning. Environmental Modelling & Software 139, 105002. URL: https://www.sciencedirect.com/science/article/pii/S1364815221000451, doi:{https://doi.org/10.1016/j.envsoft.2021.105002}.

Chen, H., Luo, Y., Potter, C., Moran, P.J., Grieneisen, M.L., Zhang, M., 2017. Modeling pesticide diuron loading from the San Joaquin watershed into the Sacramento-San Joaquin Delta using SWAT. Water Research 121, 374–385. URL: http://dx.doi.org/10.1016/j.watres.2017.05.032, doi:{10.1016/j.watres.2017.05.032}.

Chen, Z., Lin, X., Xiong, C., Chen, N., 2020. Modeling the relationship of precipitation and water level using grid precipitation products with a neural network model. Remote sensing 12, 1096. URL: http://dx.doi.org/10.3390/rs12071096, doi:{10.3390/rs12071096}.

Chen, Z., Xu, H., Jiang, P., Yu, S., Lin, G., Bychkov, I., Hmelnov, A., Ruzhnikov, G., Zhu, N., Liu, Z., 2021. A transfer Learning-Based LSTM strategy for imputing Large-Scale consecutive missing data and its application in a water quality prediction system. Journal of hydrology 602, 126573. URL: http://dx.doi.org/10.1016/j.jhydrol.2021.126573, doi:{10.1016/j.jhydrol.2021.126573}.

Cheng, X., Huang, Y., Li, R., Pu, X., Huang, W., Yuan, X., 2020. Impacts of water temperature on phosphorus release of sediments under flowing overlying water. Journal of contaminant hydrology 235, 103717. URL: http://dx.doi.org/10.1016/j.jconhyd.2020.103717, doi:{10.1016/j.jconhyd.2020.103717}.

Chivers, B.D., Wallbank, J., Cole, S.J., Sebek, O., Stanley, S., Fry, M., Leontidis, G., 2020. Imputation of missing sub-hourly precipitation data in a large sensor network: A machine learning approach. Journal of hydrology 588, 125126. URL: http://dx.doi.org/10.1016/j.jhydrol.2020.125126, doi:{10.1016/j.jhydrol.2020.125126}.

Chrobak, G., Kowalczyk, T., Fischer, T.B., Szewrański, S., Chrobak, K., Wąsowicz, B., Kazak, J.K., 2022. First, do no harm - Missing data treatment to support lake ecological condition assessment. Environmental Modelling & Software 158, 105558. URL: https://www.sciencedirect.com/science/article/pii/S1364815222002584, doi:{https://doi.org/10.1016/j.envsoft.2022.105558}.

Crisci, C., Terra, R., Pacheco, J.P., Ghattas, B., Bidegain, M., Goyenola, G., Lagomarsino, J.J., Méndez, G., Mazzeo, N., 2017. Multi-model approach to predict phytoplankton biomass and composition dynamics in a eutrophic shallow lake governed by extreme meteorological events. Ecological modelling 360, 80–93. URL: https://www.sciencedirect.com/science/article/pii/S0304380016304422, doi:{10.1016/j.ecolmodel.2017.06.017}.

Cule, E., De Iorio, M., 2012. A semi-automatic method to guide the choice of ridge parameter in ridge regression. URL: http://arxiv.org/abs/1205.0686.

Dang, X., Peng, H., Wang, X., Zhang, H., . Theil-Sen estimators in a multiple linear regression model. URL: http://www.olemiss.edu/~xdang/papers/MTSE.pdf.

[dataset] DINACEA, 2020. OAN. National Environmental Observatory. URL: https://www.ambiente.gub.uy/oan/datos-abiertos/. (Accessed on 16 November, 2022).

[dataset] Environmental data imputation project, 2022a. Metereological, Hydrometric and Data Quality variables, Santa Lucía Chico, Uruguay. URL: https://gitlab.com/fing-hydroinformatics/fsda-lu-wq/-/tree/paper/data/datasets.

[dataset] Environmental data imputation project, 2022b. Metereological, Hydrometric and Data Quality variables, Santa Lucía Chico, Uruguay. URL: https://gitlab.com/fing-hydroinformatics/fsda-lu-wq/-/tree/paper/data/imputations.

[dataset] MGAP, 2020. RENARE. Digital Terrain Model elaborate by Uruguayan National Board of Renewable Natural Resources of the Ministry of Livestock, Agriculture and Fisheries (MGAP). Available on request at Uruguay Spatial Data Infrastructure (ideuy@ide.gub.uy).

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via theEMAlgorithm. Journal of the Royal Statistical Society 39, 1–22. URL: http://dx.doi.org/10.1111/j.2517-6161.1977.tb01600.x, doi:{10.1111/j.2517-6161.1977.tb01600.x}.

Durbin, J., Koopman, S.J., 2012. Time series analysis by state space methods. OUP Oxford.

Ellison, M.E., Brett, M.T., 2006. Particulate phosphorus bioavailability as a function of stream flow and land cover. Water research 40, 1258–1268. URL: http://dx.doi.org/10.1016/j.watres.2006.01.016, doi:{10.1016/j.watres.2006.01.016}.

Fortin, M.J., Dale, M., 2005. Spatial Analysis: A Guide for Ecologists. 1 ed., Cambridge University Press, Cambridge, England.

Freni, G., Mannina, G., 2012. The identifiability analysis for setting up measuring campaigns in integrated water quality modelling. Physics and chemistry of the earth (2002) 42–44, 52–60. URL: http://dx.doi.org/10.1016/j.pce.2011.06.001, doi:{10.1016/j.pce.2011.06.001}.

Freni, G., Mannina, G., Viviani, G., 2009. Assessment of data availability influence on integrated urban drainage modeling uncertainty. Environmental Modelling & Software 24, 1171–1181. URL: https://www.sciencedirect.com/science/article/pii/S136481520900084X, doi:{https://doi.org/10.1016/j.envsoft.2009.03.007}.

Freni, G., Mannina, G., Viviani, G., 2011. Assessment of the integrated urban water quality model complexity through identifiability analysis. Water research 45, 37–50. URL: http://dx.doi.org/10.1016/j.watres.2010.08.004, doi:{10.1016/j.watres.2010.08.004}.

Geurts, P., Ernst, D., Wehenkel, L., 2006. Extremely randomized trees. Machine learning 63, 3–42. URL: http://dx.doi.org/10.1007/s10994-006-6226-1, doi:{10.1007/s10994-006-6226-1}.

Gill, M.K., Asefa, T., Kaheil, Y., McKee, M., 2007. Effect of missing data on performance of learning algorithms for hydrologic predictions: Implications to an imputation technique. Water Resources Research 43, . URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2006WR005298, doi:{https://doi.org/10.1029/2006WR005298}.

Gong, G.C., Shiah, F.K., Liu, K.K., Wen, Y.H., Liang, M.H., 2000. Spatial and temporal variation of chlorophyll a, primary productivity and chemical hydrography in the southern East China Sea. Continental shelf research 20, 411–436. URL: http://dx.doi.org/10.1016/s0278-4343(99)00079-5, doi:{10.1016/s0278-4343(99)00079-5}.

Gorgoglione, A., Bombardelli, F.A., Pitton, B.J., Oki, L.R., Haver, D.L., Young, T.M., 2019. Uncertainty in the parameterization of sediment build-up and wash-off processes in the simulation of sediment transport in urban areas. Environmental Modelling & Software 111, 170–181. URL: https://www.sciencedirect.com/science/article/pii/S1364815217307491, doi:{https://doi.org/10.1016/j.envsoft.2018.09.022}.

Gorgoglione, A., Castro, A., Chreties, C., Etcheverry, L., 2020a. Overcoming Data Scarcity in Earth Science. Data 5, 5. URL: http://dx.doi.org/10.3390/data5010005, doi:{10.3390/data5010005}.

Gorgoglione, A., Gioia, A., Iacobellis, V., 2019. A framework for assessing modeling performance and effects of rainfall-catchment-drainage characteristics on nutrient urban runoff in poorly gauged watersheds. Sustainability 11, 4933. URL: http://dx.doi.org/10.3390/su11184933, doi:{10.3390/su11184933}.

Gorgoglione, A., Gregorio, J., Ríos, A., Alonso, J., Chreties, C., Fossati, M., 2020b. Influence of land use/land cover on surface-water quality of Santa Lucía river, Uruguay. Sustainability 12, 4692. URL: http://dx.doi.org/10.3390/su12114692, doi:{10.3390/su12114692}.

Graham, J.W., 2009. Missing data analysis: making it work in the real world. Annual review of psychology 60, 549–576. URL: http://dx.doi.org/10.1146/annurev.psych.58.110405.085530, doi:{10.1146/annurev.psych.58.110405.085530}.

Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. Journal of hydrology 377, 80–91. URL: http://dx.doi.org/10.1016/j.jhydrol.2009.08.003, doi:{10.1016/j.jhydrol.2009.08.003}.

Göransson, G., Larson, M., Bendz, D., 2013. Variation in turbidity with precipitation and flow in a regulated river system – river Göta Älv, SW Sweden. Hydrology and earth system sciences 17, 2529–2542. URL: http://dx.doi.org/10.5194/hess-17-2529-2013, doi:{10.5194/hess-17-2529-2013}.

Haakonsson, S., Rodríguez, M.A., Carballo, C., Pérez, M.D.C., Arocena, R., Bonilla, S., 2020. Predicting cyanobacterial biovolume from water temperature and conductivity using a Bayesian compound Poisson-Gamma model. Water research 176, 115710. URL: http://dx.doi.org/10.1016/j.watres.2020.115710, doi:{10.1016/j.watres.2020.115710}.

Haakonsson, S., Rodríguez-Gallego, L., Somma, A., Bonilla, S., 2017. Temperature and precipitation shape the distribution of harmful cyanobacteria in subtropical lotic and lentic ecosystems. The Science of the total environment 609, 1132–1139. URL: http://dx.doi.org/10.1016/j.scitotenv.2017.07.067, doi:{10.1016/j.scitotenv.2017.07.067}.

Harris, C.R., Millman, K.J., van der Walt, S.J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M.H., Brett, M., Haldane, A., Del Río, J.F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., Oliphant, T.E., 2020. Array programming with NumPy. Nature 585, 357–362. URL: http://dx.doi.org/10.1038/s41586-020-2649-2, doi:{10.1038/s41586-020-2649-2}.

Hayashi, M., 2004. Temperature-electrical conductivity relation of water for environmental monitoring and geophysical data inversion. Environmental monitoring and assessment 96, 119–128. URL: http://dx.doi.org/10.1023/b:emas.0000031719.83065.68, doi:{10.1023/b:emas.0000031719.83065.68}.

Hoerl, A.E., Kennard, R.W., 2000. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics: a journal of statistics for the physical, chemical, and engineering sciences 42, 80. URL: http://dx.doi.org/10.2307/1271436, doi:{10.2307/1271436}.

Honaker, J., King, G., Blackwell, M., 2009. AMELIA II: A program for missing data. URL: https://r.iq.harvard.edu/docs/tmp_bak/amelia.pdf.

INIA, 2020. Uruguayan National Institute of Agricultural Research. URL: http://www.inia.uy/gras/Clima/Banco-datos-agroclimatico. (accessed on 8 November, 2022).

INUMET, 2020. Uruguayan Institute of Meteorology. URL: https://www.inumet.gub.uy/. (accessed on 10 November, 2022).

Iriarte, J.L., González, H.E., Liu, K.K., Rivas, C., Valenzuela, C., 2007. Spatial and temporal variability of chlorophyll and primary productivity in surface waters of southern Chile (41.5–43° S). Estuarine, coastal and shelf science 74, 471–480. URL: http://dx.doi.org/10.1016/j.ecss.2007.05.015, doi:{10.1016/j.ecss.2007.05.015}.

Jones, R.M., Stayner, L.T., Demirtas, H., 2014. Multiple imputation for assessment of exposures to drinking water contaminants: evaluation with the Atrazine Monitoring Program. Environmental research 134, 466–473. URL: http://dx.doi.org/10.1016/j.envres.2014.07.027, doi:{10.1016/j.envres.2014.07.027}.

Kabir, G., Tesfamariam, S., Hemsing, J., Sadiq, R., 2020. Handling incomplete and missing data in water network database using imputation methods. Sustainable and resilient infrastructure 5, 365–377. URL: http://dx.doi.org/10.1080/23789689.2019.1600960, doi:{10.1080/23789689.2019.1600960}.

Knoben, W.J.M., Freer, J.E., Woods, R.A., 2019. Technical note: Inherent benchmark or not? Comparing Nash-Sutcliffe and Kling-Gupta efficiency scores. Hydrology and earth system sciences discussions , 1–7URL: http://dx.doi.org/10.5194/hess-2019-327, doi:{10.5194/hess-2019-327}.

Kramer, O., 2013. K-Nearest Neighbors. Springer Berlin Heidelberg, Berlin, Heidelberg. p. 13–23.

Kärcher, O., Filstrup, C.T., Brauns, M., Tasevska, O., Patceva, S., Hellwig, N., Walz, A., Frank, K., Markovic, D., 2020. Chlorophyll a relationships with nutrients and temperature, and predictions for lakes across perialpine and Balkan mountain regions. Inland waters: journal of the International Society of Limnology 10, 29–41. URL: http://dx.doi.org/10.1080/20442041.2019.1689768, doi:{10.1080/20442041.2019.1689768}.

Körner, P., Kronenberg, R., Genzel, S., Bernhofer, C., 2018. Introducing Gradient Boosting as a universal gap filling tool for meteorological time series. Meteorologische Zeitschrift 27, 369–376. URL: http://dx.doi.org/10.1127/metz/2018/0908, doi:{10.1127/metz/2018/0908}.

Lintern, A., Webb, J.A., Ryu, D., Liu, S., Bende-Michl, U., Waters, D., Leahy, P., Wilson, P., Western, A.W., 2018. Key factors influencing differences in stream water quality across space: Key factors influencing differences in stream water quality across space. WIREs. Water 5, e1260. URL: http://dx.doi.org/10.1002/wat2.1260, doi:{10.1002/wat2.1260}.

McKinney, W., 2010. Data Structures for Statistical Computing in Python, in: Proceedings of the 9th Python in Science Conference, SciPy.

Mital, U., Dwivedi, D., Brown, J.B., Faybishenko, B., Painter, S.L., Steefel, C.I., 2020. Sequential imputation of missing spatio-temporal precipitation data using random forests. Frontiers in Water 2. URL: http://dx.doi.org/10.3389/frwa.2020.00020, doi:{10.3389/frwa.2020.00020}.

Moriasi, D.N., Arnold, J.G., Liew, M.W.V., Bingner, R.L., Harmel, R.D., Veith, T.L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Transactions of the ASABE 50, 885–900. URL: http://dx.doi.org/10.13031/2013.23153, doi:{10.13031/2013.23153}.

Moriasi, D.N., Gitau, M.W., Pai, N., Daggupati, P., 2015. Hydrologic and water quality models: Performance measures and evaluation criteria. Transactions of the ASABE URL: https://handle.nal.usda.gov/10113/62083.

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I — A discussion of principles. Journal of hydrology 10, 282–290. URL: http://dx.doi.org/10.1016/0022-1694(70)90255-6, doi:{10.1016/0022-1694(70)90255-6}.

Navas, R., Alonso, J., Gorgoglione, A., Vervoort, R.W., 2019. Identifying climate and human impact trends in streamflow: A case study in Uruguay. Water 11, 1433. URL: http://dx.doi.org/10.3390/w11071433, doi:{10.3390/w11071433}.

Oriani, F., Borghi, A., Straubhaar, J., Mariethoz, G., Renard, P., 2016. Missing data simulation inside flow rate time-series using multiple-point statistics. Environmental modelling & software: with environment data news 86, 264–276. URL: http://dx.doi.org/10.1016/j.envsoft.2016.10.002, doi:{10.1016/j.envsoft.2016.10.002}.

Owen, A.B., 2006. A robust hybrid of lasso and ridge regression. URL: https://statweb.stanford.edu/~owen/reports/hhu.pdf.

Paaijmans, K.P., Takken, W., Githeko, A.K., Jacobs, A.F.G., 2008. The effect of water turbidity on the near-surface water temperature of larval habitats of the malaria mosquito Anopheles gambiae. International journal of biometeorology 52, 747–753. URL: http://dx.doi.org/10.1007/s00484-008-0167-2, doi:{10.1007/s00484-008-0167-2}.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2012. Scikit-learn: Machine Learning in Python. URL: http://arxiv.org/abs/1201.0490.

Poll, W.H., Maat, D.S., Fischer, P., Visser, R.J.W., Brussaard, C.P.D., Buma, A.G.J., 2021. Solar radiation and solar radiation driven cycles in warming and freshwater discharge control seasonal and inter-annual phytoplankton chlorophyll a and taxonomic composition in a high Arctic fjord (Kongsfjorden, Spitsbergen). Limnology and oceanography 66, 1221–1236. URL: http://dx.doi.org/10.1002/lno.11677, doi:{10.1002/lno.11677}.

Ratolojanahary, R., Houé Ngouna, R., Medjaher, K., Junca-Bourié, J., Dauriac, F., Sebilo, M., 2019. Model selection to improve multiple imputation for handling high rate missingness in a water quality dataset. Expert systems with applications 131, 299–307. URL: http://dx.doi.org/10.1016/j.eswa.2019.04.049, doi:{10.1016/j.eswa.2019.04.049}.

Rodríguez, R., Pastorini, M., Etcheverry, L., Chreties, C., Fossati, M., Castro, A., Gorgoglione, A., 2021. Water-quality data imputation with a high percentage of missing values: A machine learning approach. Sustainability 13, 6318. URL: http://dx.doi.org/10.3390/su13116318, doi:{10.3390/su13116318}.

Ríos, A., 2019. Implementación de un modelo hidrodinámico tridimensional en el embalse de Paso Severino. Aportes para la modelación de calidad de agua. Ph.D. thesis. Universidad de la República. Facultad de Ingeniería.. Montevideo, Uruguay. URL: https://www.colibri.udelar.edu.uy/jspui/handle/20.500.12008/21553.

Saalidong, B.M., Aram, S.A., Otu, S., Lartey, P.O., 2022. Examining the dynamics of the relationship between water pH and other water quality parameters in ground and surface water systems. PloS one 17, e0262117. URL: http://dx.doi.org/10.1371/journal.pone.0262117, doi:{10.1371/journal.pone.0262117}.

Satpathy, K.K., Mohanty, A.K., Sahu, G., Sarguru, S., Sarkar, S.K., Natesan, U., 2011. Spatio-temporal variation in physicochemical properties of coastal waters off Kalpakkam, southeast coast of India, during summer, pre-monsoon and post-monsoon period. Environmental monitoring and assessment 180, 41–62. URL: http://dx.doi.org/10.1007/s10661-010-1771-2, doi:{10.1007/s10661-010-1771-2}.

Sattari, M.T., Rezazadeh-Joudi, A., Kusiak, A., 2017. Assessment of different methods for estimation of missing data in precipitation studies. Hydrology research 48, 1032–1044. URL: http://dx.doi.org/10.2166/nh.2016.364, doi:{10.2166/nh.2016.364}.

Shinohara, R., Tanaka, Y., Kanno, A., Matsushige, K., 2021. Relative impacts of increases of solar radiation and air temperature on the temperature of surface water in a shallow, eutrophic lake. Hydrology research 52, 916–926. URL: http://dx.doi.org/10.2166/nh.2021.148, doi:{10.2166/nh.2021.148}.

Song, J.H., Her, Y., Guo, T., 2022. Quantifying the contribution of direct runoff and baseflow to nitrogen loading in the Western Lake Erie Basins. Scientific reports 12, 9216. URL: http://dx.doi.org/10.1038/s41598-022-12740-1, doi:{10.1038/s41598-022-12740-1}.

Stockman, M., Dwivedi, D., Gentz, R., Peisert, S., 2019. Detecting control system misbehavior by fingerprinting programmable logic controller functionality. International journal of critical infrastructure protection 26, 100306. URL: http://dx.doi.org/10.1016/j.ijcip.2019.100306, doi:{10.1016/j.ijcip.2019.100306}.

Suykens, J.A.K., Vandewalle, J., 1999. Least Squares Support Vector Machine Classifiers. Neural processing letters 9, 293–300. URL: http://dx.doi.org/10.1023/a:1018628609742, doi:{10.1023/a:1018628609742}.

Tabari, H., Hosseinzadeh Talaee, P., 2015. Reconstruction of river water quality missing data using artificial neural networks. Water Quality Research Journal 50, 326–335. URL: http://dx.doi.org/10.2166/wqrjc.2015.044, doi:{10.2166/wqrjc.2015.044}.

Templ, M., Kowarik, A., Filzmoser, P., 2011. Iterative stepwise regression imputation using standard and robust methods. Computational statistics & data analysis 55, 2793–2806. URL: http://dx.doi.org/10.1016/j.csda.2011.04.012, doi:{10.1016/j.csda.2011.04.012}.

Tencaliec, P., Favre, A.C., Prieur, C., Mathevet, T., 2015. Reconstruction of missing daily streamflow data using dynamic regression models. Water Resources Research 51, 9447–9463. URL: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1002/2015WR017399, doi:{https://doi.org/10.1002/2015WR017399}.

Tipping, M.E., 2001. Sparse Bayesian learning and the relevance vector machine. Journal of machine learning research 1, 211–244.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R.B., 2001. Missing value estimation methods for DNA microarrays. Bioinformatics (Oxford, England) 17, 520–525. URL: http://dx.doi.org/10.1093/bioinformatics/17.6.520, doi:{10.1093/bioinformatics/17.6.520}.

Vilaseca, F., Castro, A., Chreties, C., Gorgoglione, A., 2023. Assessing influential rainfall–runoff variables to simulate daily streamflow using random forest. Hydrological Sciences Journal 68, 1738–1753. doi:10.1080/02626667.2023.2232356.

Villa, A., Fölster, J., Kyllmar, K., 2019. Determining suspended solids and total phosphorus from turbidity: comparison of high-frequency sampling with conventional monitoring methods. Environmental monitoring and assessment 191, 605. URL: http://dx.doi.org/10.1007/s10661-019-7775-7, doi:{10.1007/s10661-019-7775-7}.

Villate, F., Aravena, G., Iriarte, A., Uriarte, I., 2008. Axial variability in the relationship of chlorophyll a with climatic factors and the North Atlantic Oscillation in a Basque coast estuary, Bay of Biscay (1997-2006). Journal of plankton research 30, 1041–1049. URL: http://dx.doi.org/10.1093/plankt/fbn056, doi:{10.1093/plankt/fbn056}.

Wang, X., Li, A., Jiang, Z., Feng, H., 2006. Missing value estimation for DNA microarray gene expression data by Support Vector Regression imputation and orthogonal coding scheme. BMC bioinformatics 7, 32. URL: http://dx.doi.org/10.1186/1471-2105-7-32, doi:{10.1186/1471-2105-7-32}.

White, I.R., Carlin, J.B., 2010. Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Statistics in medicine 29, 2920–2931. URL: http://dx.doi.org/10.1002/sim.3944, doi:{10.1002/sim.3944}.

Wolpert, D.H., Macready, W.G., 1997. No free lunch theorems for optimization. IEEE transactions on evolutionary computation: a publication of the IEEE Neural Networks Council 1, 67–82. URL: http://dx.doi.org/10.1109/4235.585893, doi:{10.1109/4235.585893}.

Ye, X., Xu, C.Y., Li, Y., Li, X., Zhang, Q., 2017. Change of annual extreme water levels and correlation with river discharges in the middle-lower Yangtze River: Characteristics and possible affecting factors. Chinese geographical science 27, 325–336. URL: http://dx.doi.org/10.1007/s11769-017-0866-x, doi:{10.1007/s11769-017-0866-x}.

Zhang, H., Wang, L., 2021. Analysis of the variation in potential evapotranspiration and surface wet conditions in the Hancang River Basin, China. Scientific reports 11, 8607. URL: http://dx.doi.org/10.1038/s41598-021-88162-2, doi:{10.1038/s41598-021-88162-2}.

Zhang, Y., Thorburn, P.J., 2021. A dual-head attention model for time series data imputation. Computers and Electronics in Agriculture 189, 106377. URL: https://www.sciencedirect.com/science/article/pii/S016816992100394X, doi:https://doi.org/10.1016/j.compag.2021.106377.

Zhang, Y., Thorburn, P.J., 2022. Handling missing data in near real-time environmental monitoring: A system and a review of selected methods. Future generations computer systems: FGCS 128, 63–72. URL: https://www.sciencedirect.com/science/article/pii/S0167739X21003794, doi:{10.1016/j.future.2021.09.033}.