 doi:10.20944/preprints202105.0105.v1

Article

Water-quality data imputation with high percentage of missing values: a machine learning approach

Rafael Rodriguez[1], Marcos Pastorini[2], Lorena Etcheverry[2], Christian Chreties[1], Mónica Fossati[1], Alberto Castro[2] and Angela Gorgoglione[1,*]

[1] Department of Fluid Mechanics and Environmental Engineering (IMFIA), School of Engineering, Universidad de la República, Montevideo 11300, Uruguay; rrodriguez@fing.edu.uy, chreties@fing.edu.uy, mfossati@fing.edu.uy, agorgoglione@fing.edu.uy.

[2] Department of Computer Science (InCo), School of Engineering, Universidad de la República, Montevideo 11300, Uruguay; mpastorini@fing.edu.uy, lorenae@fing.edu.uy, acastro@fing.edu.uy.

* Correspondence: agorgoglione@fing.edu.uy

Abstract: The monitoring of surface-water quality followed by water-quality modeling and analysis is essential for generating effective strategies in water-resource management. However, worldwide, particularly in developing countries, water-quality studies are limited due to the lack of a complete and reliable dataset of surface-water-quality variables. In this context, several statistical and machine-learning models were assessed for imputing water-quality data at six monitoring stations located in the Santa Lucía Chico river (Uruguay), a mixed lotic and lentic river system. The challenge of this study is represented by the high percentage of missing data (between 50% and 70%) and the high temporal and spatial variability that characterizes the water-quality variables. The competing algorithms implemented belonged to both univariate and multivariate imputation methods (inverse distance weighting (IDW), Random Forest Regressor (RFR), Ridge (R), Bayesian Ridge (BR), AdaBoost (AB), Hubber Regressor (HR), Support Vector Regressor (SVR), and K-nearest neighbors Regressor (KNNR)). According to the results, more than 76% of the imputation outcomes are considered satisfactory (NSE > 0.45). The imputation performance shows better results at the monitoring stations located inside the reservoir than the ones positioned along the mainstream. IDW was the most chosen model for data imputation.

Keywords: data scarcity; water quality; missing data; univariate imputation; multivariate imputation; machine learning; hydroinformatics.

1. Introduction

Monitoring, modeling, and management represent the three foundations for building an effective pollution-control strategy [1]. They strictly depend on each other: there is no management without modeling and no modeling without exhaustive monitoring. Therefore, any problem related to data collection is then reflected in the performance of the modeling and management phases. Consequently, it is crucial first to acknowledge what improvement would result if all the available data could be well exploited [2].

The issue of missing data frequently occurs in environmental fields due to sensor failures, weak or inexistent strategy for coordinating monitoring campaigns, a change in the measurement site, in data collectors, or to the equipment over time, budget issues [3, 4]. This water-quality series problem is particularly significant in developing countries where monitoring stations and/or monitoring frequency is scarce, and the percentage of missing data is exceptionally high [5].

It is possible to deal with missing data in two different ways: deletion or imputation [6]. Deletion consists of removing the observations or the features characterized by missing values, while imputation involves reconstructing missing data. Deletion is typically the default method adopted since it is rapid and straightforward [7]. However, in several fields, there are many examples in which such technique presented some restrictions. In

 doi:10.20944/preprints202105.0105.v1

fact, it reduces the dataset size and may lead to bias results and a loss of critical information, mainly when a high percentage of missing values characterizes the dataset. Among the most straightforward imputation techniques, there are mean imputation and linear interpolation (that just depend on the time series' available data to be imputed), arithmetic and weighted averaging. However, these techniques have shown poor performance when the dataset is characterized by a significant length of the missing sequence [5, 8].

Another common approach used to complete missing data, which is part of the univariate imputation methods, is to adopt information from the neighboring monitoring stations. The inverse distance weight (IDW) is one of those techniques that has been successfully adopted for environmental datasets, particularly for meteorological variables [9-11].

In the last decade, progressively more advanced techniques have been adopted to reconstruct environmental time series. Among them, the machine learning ones that can handle multivariate inputs are the most extensively used. Aguilera et al. [5] adopted three different methods (spatio-temporal kriging, multiple imputations by chained equations through predictive mean matching, and random forest) to reconstruct daily precipitation time series characterized by extreme missingness (>90%). They found that spatio-temporal kriging simulates rainfall distribution under missing chronological patterns more reliably than the other two techniques adopted. Sattari et al. [12] provided an in-depth comparison of ten different statistical and machine-learning models to impute monthly precipitation data. The computational outcomes demonstrated that among the classical statistical methods, arithmetic averaging, multiple linear regressor, and non-linear iterative partial least squares perform better. The multiple imputation technique performed better when rainfall data from more than one dependent station were considered. Also Barrios et al. [10] compared the performance of five models to infill monthly precipitation records, finding that artificial neural network, multiple linear regression, and IDW showed the best performance.

It is clear that most of the imputation works presented in the scientific literature refer to meteorological variables and, sometimes to streamflow (hydrologic variables) [13]. To our knowledge, there are few works related to water-quality data imputation. Among them, Tabari and Talaee [14] adopted artificial neural networks to successfully recover the missing values of 13 water-quality parameters at five monitoring stations in the South of Iran. Srebotnjak et al. [15] adopted hot-deck imputation to improve a country-level water quality index, calculated considering dissolved oxygen, electrical conductivity, *pH*, total phosphorus, and total nitrogen. Ratolojanahary et al. [16] tackled for the very first time the problem of high rate missingness in a water-quality dataset by adopting four machine learning models (random forest, boosted regression trees, k-nearest neighbors, and support vector regression). However, there is not an exhaustive evaluation of different types of imputation models in the context of water-quality data characterized by a high percentage of incompleteness.

Since the start of systematic water-quality monitoring in 2004, Uruguay has been suffering the problem of data scarcity, causing significant limitations in the development and implementation of reliable and accurate water-quality models. This unavoidably produces the lack of management tools to design effective policies aimed at mitigating pollution impacts on receiving water bodies.

Based on these considerations, this study aims at augmenting the current water-quality dataset of one of the most important Uruguayan watersheds, Santa Lucía Chico. In particular, we assess the performance of several univariate and multivariate imputation models (statistical and machine learning) to impute missing bi-monthly water-quality data (water temperature, electrical conductivity, pH, dissolved oxygen, total nitrogen, nitrite, nitrate, and turbidity) and double them. The challenge of this work is represented by the high missingness percentage (between 50% and 70%) and the high temporal and spatial distribution of the variables under study.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 May 2021 doi:10.20944/preprints202105.0105.v1

2. Materials and Methods

2.1. Dataset description

The water-quality dataset used in this study includes the following physical and chemical variables: water temperature (T_w) [ºC], electrical conductivity (*EC*) [µS/cm], pH, dissolved oxygen (*DO*) [mg/L], total nitrogen (*TN*) [mg/L], nitrite (NO_2^-) [mg/L], nitrate (NO_3^-) [mg/L], and turbidity (*Turb*) [NTU]. It was recorded by the Uruguayan National Environment Board (DINAMA) and is freely downloadable from the National Environmental Observatory (OAN) [17]. Data were collected from 2014 to 2020, with a bi-monthly frequency, at six monitoring stations located along the Santa Lucía Chico river, South America, Uruguay. It is a mixed lotic and lentic system with wise national importance since its waters flow into the Paso Severino reservoir, the primary national drinking water source [18, 19, 20, 21]. The first three upstream monitoring stations (SLC01, SLC02, and PS01) are located before the reservoir; the other three stations (PS03, PS04, and PS02) are located in the lake (Figure 1).

The percentage of missing values detected for each variable at all monitoring stations is reported in Table 1.

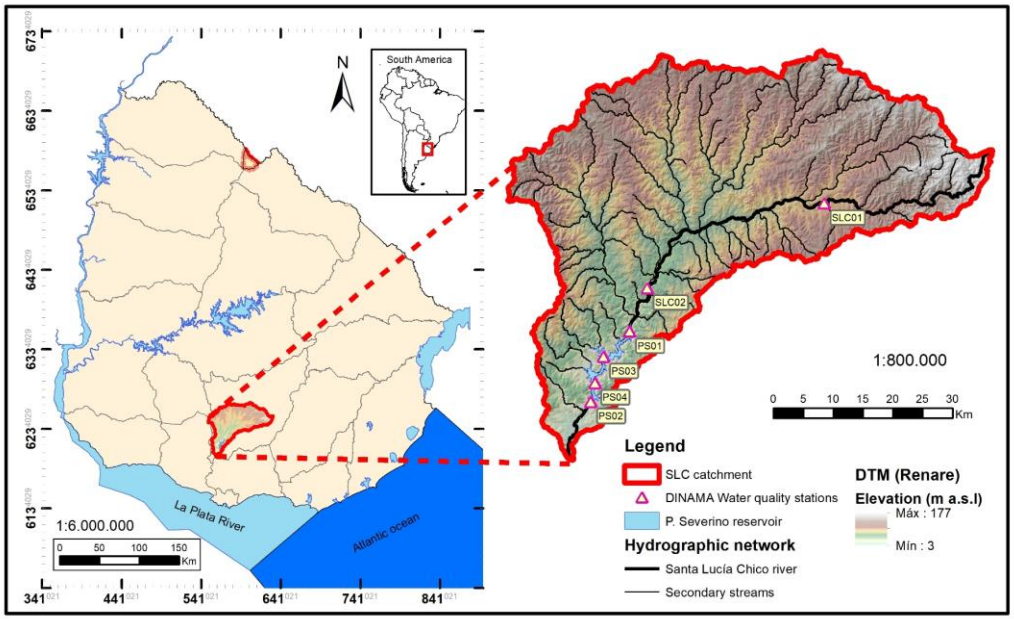

Figure 1. Santa Lucía Chico river (Uruguay) and location of the six water-quality monitoring stations.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 May 2021 doi:10.20944/preprints202105.0105.v1

Table 1. Percentage of missing data for the variables under study at the six monitoring stations (period 2014-2020).

Variable		% missing data					
		SLC01	SLC02	PS01	PS03	PS04	PS02
Physical	T_w	51.5	51.5	64.7	57.6	57.6	59.1
	EC	51.5	51.5	64.7	57.6	57.6	57.6
	pH	52.9	52.9	66.2	59.1	59.1	59.1
	DO	51.5	51.5	64.7	57.6	57.6	57.6
	Turb	52.9	52.9	66.2	60.6	60.6	59.1
Chemical	*TN*	52.9	52.9	66.2	60.6	60.6	59.1
	NO_2^-	51.5	51.5	64.7	59.1	59.1	57.6
	NO_3^-	51.5	51.5	64.7	59.1	59.1	57.6

As support for multivariate techniques, hydro-meteorological variables that may influence the water-quality variables under study were also considered (period 2014-2020). In particular, air temperature (T_a) (minimum, average, maximum) [°C], solar radiation (*SR*) [cal/cm²/d], and heliophany (*Hel*) (sunshine hours) [h] were taken into account for the imputation of T_w. They were daily collected by the National Institute of Agricultural Research (INIA) and they do not have any missing data.

T_a along with daily evapotranspiration (ET) data, also calculated from INIA (time series characterized by 0.1% of missing data), were considered for the imputation of *Turb*.

Streamflow (*Q*) [m³/s] was considered for the imputation of *TN*, NO_2^-, NO_3^-, and *Turb*. This time series was measured three times a day by the Uruguay National Water Board (DINAGUA) and is characterized by a neglectable percentage of missing data (5.6%).

Furthermore, precipitation records (P) from the Uruguayan Institute of Meteorology (INUMET) were considered for *Turb* imputation. The time series observed at the ten selected monitoring stations have a percentage of missing data that varies between 0.0% and 8.6% in the considered time window (2014-2020). The location of the INIA, DINAGUA, and INUMET monitoring stations is represented in Figure 2.

Figure 2. Location of the meteorological (INIA), hydrometric (DINAGUA), and pluviometric (INUMET) monitoring stations in Santa Lucía Chico.

2.2. Imputation techniques

Since the best model for imputing any kind of variable does not exist [22], we adopted several statistical and machine-learning algorithms (single and multiple imputation) to accomplish the objective of this study. The selected models are Inverse distance weighting

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 May 2021 doi:10.20944/preprints202105.0105.v1

(IDW), Random Forest regressor (RFR), Ridge regressor (RR), Bayesian ridge (BR), AdaBoost (AB), Huber regressor (HR), Support vector regressor (SVR), TheilSen regressor (TSR), and k-nearest neighbors regressor (KNNR). All of them have proved to be suited for non-linear environmental variables, and some of them for cases characterized by a high percentage of missing data. Furthermore, they are already programmed and freely available in Phyton. Unless a software library is explicitly mentioned, *scikit-learn* was adopted to implement the algorithms [23].

Inverse Distance Weighting (*IDW*). It is a deterministic method for univariate interpolation. Missing samples from the target station (*s*) are computed from the values observed in the neighboring stations. Weighting is assigned to data using a weighting power that controls how the weighting factors drop off as the distance from the station *s* increases [24].

Random Forest Regressor (*RFR*). It is a supervised learning algorithm that uses an ensemble learning method for regression. Such method is a technique that combines predictions from multiple Decision Tree algorithms to improve the overall prediction and control overfitting. The decision trees run in parallel with no interaction among them and the mean of all the predictions is returned [25].

Ridge Regressor (*RR*). It is a technique for analyzing multiple regressions of highly correlated data. It trains a regression model that seeks to minimize the least-squares function with an additional regularization term given by the sum of the values' squares (L2 norm) [26].

Bayesian Ridge (BR). It is an estimator that assumes and predicts the target calculating its probability distribution during training. This method can overcome data sparsity more correctly than other methods [27].

AdaBoost (AB). It is an estimator that starts fitting a Decision Tree regressor on the original dataset and then fits additional copies of the regressor on the same dataset slightly modified. Depending on the correctness of the last prediction, samples that are difficult to predict become more relevant as the training continues. The mean of all the models' predictions is returned [28].

Huber Regressor (HR). It is an algorithm that trains a linear model which optimizes the mean squared error (L2 error) for samples whose error is lower than a given threshold (*d*) and the mean absolute error (L1 error) for samples whose error is greater than *d*. In this way, the optimized function is not heavily influenced by outliers while not completely ignoring their effect [29].

Support Vector Regressor (SVR). It is an estimator that focuses on minimizing the coefficients, more specifically, the l2-norm of the coefficient vector, not the squared error. The error term is instead handled in the constraints, where the absolute error is set to less than or equal to a specified margin (maximum error). The latter can be tuned to gain the desired accuracy of the model [30].

TheilSen Regressor (TSR). It is a regressor that makes its estimation by calculating the slopes and intercepts of a subpopulation of all possible combinations of some subsample points. The final slope and intercept are then defined as the spatial median of these slopes and intercepts. It is robust against outliers compared to other linear regressors [31].

K-Nearest Neighbors Regressor (KNNR). It is a regressor that calculates the distance (using all variables) from the target point to the others and makes a prediction by interpolating the nearest neighbors in the dataset [32].

2.3. Imputation performance evaluation

To compare the accuracy of the implemented techniques in reconstructing missing water-quality data, the Kling-Gupta efficiency (KGE), the percent bias (PBIAS), and the Nash-Sutcliffe efficiency (NSE) were adopted (Eq. 1-Eq. 3). The latter was used as an objective function since it is the most restrictive one [33], while KGE and PBIAS were adopted for validation.

$$NSE = 1 - \frac{\sum_{i=1}^{n}(x_i^o - x_i^c)^2}{\sum_{i=1}^{n}(x_i^o - \overline{x^o})^2} \quad (1)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 May 2021 doi:10.20944/preprints202105.0105.v1

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

$$PBIAS = 100 \times \frac{\sum_{i=1}^{n}(x_i^o - x_i^c)}{\sum_{i=1}^{n}(x_i^o)} \quad (3)$$

Where x_i^o is the i^{th} observed value, x_i^c is the i^{th} computed (or imputed) value, $\overline{x^o}$ is the mean of observed values, and n is the testing-dataset size. Being (μ^c,σ^c) and (μ^o,σ^o) the first two statistical moments (mean and standard deviation) of x^c and x^o respectively, r is the linear correlation between observations and imputations, α is a measure of the flow variability error ($\alpha = \sigma^c/\sigma^o$), β is a bias term ($\beta = \mu^c/\mu^o$).

NSE varies between -∞ and 1. If NSE is equal to 1, imputed values perfectly match the records. If NSE is equal to 0, imputed values are as good as the observation mean. If NSE is negative, the observation mean is a better predictor than imputed values. Therefore, higher NSE values are desirable since they indicate a more accurate imputation model [34, 35].

Unlike NSE, there are not well-defined accepted threshold of KGE that define a "good" model. For this reason, there is a tendency in current literature to interpret KGE values in a similar way to NSE: negative values indicate "bad" model performance, whereas positive values indicate "good" model performance [36, 37, 38]. However, a recent study by Knoben et al. [39] found that all model outcomes with -0.41 < KGE < 1 could be considered as good performance.

The optimal PBIAS value is 0, with lower values indicating accurate model imputation. Positive values point out model underestimation bias, and negative values indicate model overestimation bias [40].

For this study, for NSE, KGE, and PBIAS, the corresponding performance evaluation criteria were established according to a recent review [34, 39, 40] (Table 2).

Table 2. Evaluation metrics and associated performance ratings.

Performance rating	Physical water quality variables	Chemical water quality variables
NSE		
Very good	$NSE > 0.80$	$NSE > 0.65$
Good	$0.70 < NSE \leq 0.80$	$0.50 < NSE \leq 0.65$
Satisfactory	$0.45 < NSE \leq 0.70$	$0.35 < NSE \leq 0.50$
Unsatisfactory	$NSE \leq 0.45$	$NSE \leq 0.35$
PBIAS		
Very good	$\lvert PBIAS \rvert < 10$	$\lvert PBIAS \rvert < 15$
Good	$10 \leq \lvert PBIAS \rvert < 15$	$15 \leq \lvert PBIAS \rvert < 20$
Satisfactory	$15 \leq \lvert PBIAS \rvert < 20$	$20 \leq \lvert PBIAS \rvert < 30$
Unsatisfactory	$\lvert PBIAS \rvert \geq 20$	$\lvert PBIAS \rvert \geq 30$
KGE		
Satisfactory/Good	$KGE \geq -0.41$	$KGE \geq -0.41$
Unsatisfactory	$KGE < -0.41$	$KGE < -0.41$

2.4. Aiding variables for the imputation process

Considering the correlations among water-quality variables, multivariate techniques exploited them for completing the missing values with the other existing water-quality observations. In particular, we considered that *Tw* and *Turb* influence *EC* in surface waters. An increase in *Tw* causes an increase in the mobility of the ions present in the water. An increase in *Tw* may also produce an increment in the number of ions due to molecule dissociation. As the *EC* depends on these factors, an increase in *Tw* leads to an increase in

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 May 2021 doi:10.20944/preprints202105.0105.v1

EC [41, 42]. Furthermore, *EC* represents the ability of a liquid to conduct an electric charge; this ability depends on dissolved ion concentration, which is usually measured as total dissolved solids (*TDS*) [43]. Considering that *TDS* are highly correlated with *Turb*, we can assume that *EC* is also affected by *Turb*.

Other correlations considered were the ones between *TN* - *Turb*, *TN* – NO_2^-, and *TN* - NO_3^-. This is justified by the fact that *TN* represents the sum of dissolved and particle-bound nitrogen. Furthermore, *DO* was considered dependent on T_w: the higher T_w, the lower *DO* [20].

Moreover, as we have already mentioned in section 2.1., hydro-meteorological variables were also considered as an aid for the imputation process since they may influence the water-quality variables under study. Particularly, *Tw* is considered to be mainly affected by *Ta*, *Hel*, and *SR*. *Turb*, *TN*, NO_2^-, and NO_3^- are influenced by *Q*. Considering that NO_2^- and NO_3^- are part of the dissolved inorganic nitrogen (DIN), their correlation with streamflow is clear: the higher *Q*, the lower the ions concentration, due to dilution process [20]. Being *TN* the sum of dissolved and particle-bound nitrogen, we aided the imputation process of the latter by including *Turb* data. Considering the importance of overland runoff in transporting sediments, it is often assumed that these constituents have a positive relationship to river discharge [44]. For this reason, we are considering *Q* as a supporting variable for *Turb* imputation.

In their study carried out in Santa Lucía Chico watershed, Gorgoglione et al. [20] found a seasonality of *Turb* values with higher values during the cold season. This was justified by the fact that in this season, frequent extreme precipitation events occur, and, along with higher soil humidity due to low temperature, this causes a higher runoff and, therefore, a more significant amount of detached and washed-off sediments. For this reason, *Turb* imputation is also aided by *ET*, *P*, and T_a data (apart from *Q* as previously explained).

A summary of the supporting variables taken into account for the imputation process is represented in Table 3.

Table 3. Aiding variables considered in the imputation process.

Imputing variable	Aiding variable
Water temperature (T_w)	Air temperature (T_a)
	Solar radiation (*SR*)
	Heliophany (*Hel*)
Dissolved oxygen (*DO*)	Water temperature (T_w)
Nitrite (NO_2^-)	Streamflow (*Q*)
Nitrate (NO_3^-)	Streamflow (*Q*)
Turbidity (*Turb*)	Streamflow (*Q*)
	Precipitation (*P*)
	Air temperature (T_a)
	Evapotranspiration (*ET*)
Total Nitrogen (*TN*)	Nitrite (NO_2^-)
	Nitrate (NO_3^-)
	Turbidity (*Turb*)
	Streamflow (*Q*)

3. Results and Discussion

3.1. Dataset profiling

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 May 2021 doi:10.20944/preprints202105.0105.v1

The dataset considered for this study is formed by 48 time series (8 water-quality variables × 6 monitoring stations). Therefore, from now on, we will call "variable," "feature," or "attribute," a particular time series that refers to a water-quality variable recorded at one monitoring station (e.g., T_w observed at SLC01 monitoring station will be $T_w[SLC01]$).

The data profiling process was programmed and run in Python 3.8, using the *pandas_profiling* library [45].

With the aim of showing the high temporal and spatial variability of the water-quality variables under study, we reported the box-plot representation at the six monitoring stations through the period analyzed (2014-2020) (Figure 3). From all the pollutant plots presented, it is interesting to identify two different groups of behavior: the three monitoring sites located in the reservoir show different patterns compared to the ones that characterize the stations located upstream of the reservoir. Furthermore, T_w and *DO* are the only pollutants showing a strong intra- and inter-annual seasonality, while a clear pattern cannot be identified for the other contaminants under study. It is essential to highlight the high nutrient contribution of PS01 (*TN*, NO_2^-, NO_3^-), where the biggest city of the watershed is located (Florida). It is known that urbanized areas are sources of nitrogen due to atmospheric deposition, lawn fertilizer application, wastewater effluent, and leaky sewage infrastructure [46]. *Turb* shows a minor temporal pattern through the years with the highest values registered in the monitoring stations located upstream of the reservoir (SLC01, SLC02, and PS01).

3.2. Imputation results

The implementation of the methodology adopted in this study was performed using Python programming language on a computer with the following main features: Ubuntu Operating System, 16GB of RAM, and Intel i3 Processor.

For dealing with the different units and orders of magnitude, the dataset was min-max normalized prior to any analysis.

To evaluate the performance of the different models adopted and to choose the best one for each feature, k-fold cross-validation with k=10 was used in this study. If a time series was characterized by less than 100 records, we adopted a repeated k-fold cross-validation, always with k=10. This method repeats the k-fold cross-validation process multiple times and reports the mean performance across all folds and all repeats [47]. The winning (best) models were the ones with the optimal hyper-parameters, i.e., those with the highest NSE (objective function). As a result, the best model with the highest accuracy was selected for each feature and validated by calculating KGE and PBIAS. The outcomes of this methodology are represented by augmented time series for all the water-quality variables, characterized by one-month frequency (i.e., the frequency was doubled up).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 May 2021 doi:10.20944/preprints202105.0105.v1

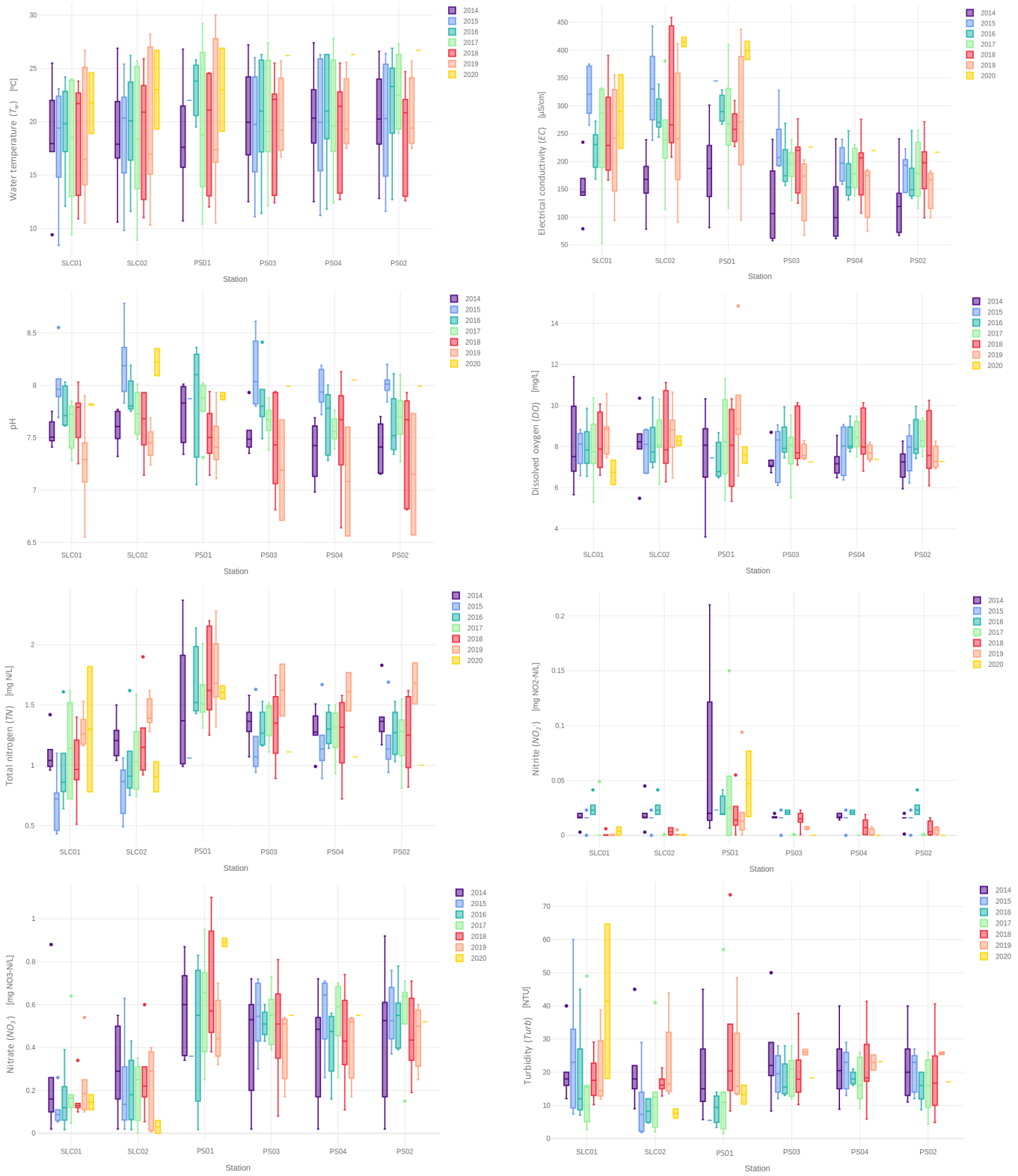

Figure 3. Temporal and spatial variation of the water-quality variables.

In Table 4, we report, for each variable, the winning model with the corresponding values of the goodness-of-fit indicators calculated and the corresponding rating based on Table 3.

Table 4. Best imputation models and corresponding goodness-of-fit indicator values per variable.

Variable	Station	Model	NSE	NSE rating	PBIAS	PBIAS rating	KGE	KGE rating

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 May 2021 doi:10.20944/preprints202105.0105.v1

T_w	SLC01	Random Forest Regressor	0.95	Very good	0.09	Very good	0.91	Good
	SLC02	IDW	0.97	Very good	-2.54	Very good	0.95	Good
	PS01	IDW	0.95	Very good	-3.77	Very good	0.94	Good
	PS03	IDW	0.98	Very good	-0.21	Very good	0.96	Good
	PS04	IDW	0.98	Very good	1.49	Very good	0.96	Good
	PS02	IDW	0.97	Very good	0.89	Very good	0.93	Good
EC	SLC01	SVR	0.67	Satisfactory	-0.12	Very good	0.76	Good
	SLC02	SVR	0.71	Good	0.43	Very good	0.67	Good
	PS01	Ridge	0.67	Satisfactory	-1.70	Very good	0.77	Good
	PS03	Ridge	0.85	Very good	1.35	Very good	0.86	Good
	PS04	IDW	0.94	Very good	4.71	Very good	0.87	Good
	PS02	IDW	0.89	Very good	-3.89	Very good	0.88	Good
pH	SLC01	Bayesian Ridge	0.39	Unsatisfactory	-0.63	Very good	0.54	Good
	SLC02	Random Forest Regressor	0.75	Good	0.95	Very good	0.80	Good
	PS01	Random Forest Regressor	0.25	Unsatisfactory	0.44	Very good	0.40	Good
	PS03	Bayesian Ridge	0.66	Satisfactory	-0.31	Very good	0.78	Good
	PS04	IDW	0.68	Satisfactory	-1.10	Very good	0.79	Good
	PS02	Huber Regressor	0.65	Satisfactory	-3.29	Very good	0.77	Good
DO	SLC01	Bayesian Ridge	0.81	Very good	-2.79	Very good	0.83	Good
	SLC02	Random Forest Regressor	0.73	Good	-1.80	Very good	0.73	Good
	PS01	AdaBoost	0.27	Unsatisfactory	-1.65	Very good	0.48	Good
	PS03	Ridge	0.80	Good	-0.15	Very good	0.86	Good
	PS04	Huber Regressor	0.89	Very good	-0.28	Very good	0.89	Good
	PS02	IDW	0.69	Satisfactory	-0.24	Very good	0.79	Good
TN	SLC01	IDW	0.19	Unsatisfactory	2.72	Very good	0.49	Good
	SLC02	Ridge	0.65	Good	1.90	Very good	0.72	Good
	PS01	Random Forest Regressor	-0.35	Unsatisfactory	-0.91	Very good	-0.10	Good
	PS03	IDW	0.63	Good	-7.79	Very good	0.75	Good
	PS04	Random Forest Regressor	0.77	Very good	-1.38	Very good	0.71	Good
	PS02	IDW	0.70	Very good	-15.22	Good	0.71	Good
NO_2^-	SLC01	Huber Regressor	0.59	Good	-0.83	Very good	0.62	Good
	SLC02	Random Forest Regressor	0.36	Satisfactory	-10.79	Very good	0.54	Good
	PS01	KNN	-0.31	Unsatisfactory	25.94	Satisfactory	0.02	Good
	PS03	TheilSen Regressor	0.74	Very good	1.09	Very good	0.72	Good
	PS04	KNN	0.92	Very good	3.35	Very good	0.86	Good
	PS02	Huber Regressor	0.75	Very good	-4.53	Very good	0.78	Good
NO_3^-	SLC01	TheilSen Regressor	0.21	Unsatisfactory	13.68	Very good	0.33	Good
	SLC02	Huber Regressor	0.42	Satisfactory	-4.95	Very good	0.58	Good
	PS01	Random Forest Regressor	0.10	Unsatisfactory	5.14	Very good	0.36	Good
	PS03	IDW	0.69	Very good	-0.80	Very good	0.80	Good
	PS04	Huber Regressor	0.80	Very good	-1.08	Very good	0.84	Good
	PS02	SVR	0.61	Good	-1.57	Very good	0.75	Good
Turb	SLC01	SVR	-0.10	Unsatisfactory	-1.93	Very good	0.03	Good

 doi:10.20944/preprints202105.0105.v1

SLC02	SVR	0.56	Satisfactory	-5.74	Very good	0.67	Good
PS01	IDW	-0.18	Unsatisfactory	-45.97	Unsatisfactory	0.35	Good
PS03	IDW	0.66	Satisfactory	-12.30	Good	0.71	Good
PS04	IDW	0.85	Very good	3.94	Very good	0.88	Good
PS02	IDW	0.88	Very good	-3.27	Very good	0.87	Good

Considering the NSE rating, the imputation performance is overall adequate. *Tw* at the six monitoring stations was the best-imputed variable, showing "very good" performance. The strong daily and annual seasonality that characterizes this variable makes its simulation easy and, therefore, its imputation. The correlation that exists between *Tw* and *EC* (an increase in *Tw* leads to an increase in *EC*) [41, 42] is reflected in the "good" performance of this variable at the six monitoring sites ("satisfactory at SLC01 and PS01; "good" at SLC02; "very good" at PS03, PS04, and PS02). The imputation process for the other water-quality variables shows different results. It is noteworthy that the performance is always notable at the three monitoring stations located in the reservoir of Paso Severino (PS03, PS04, and PS02); while the imputation can sometimes be "unsatisfactory" at the stations located upstream, along Santa Lucía Chico river (SLC01, SLC02, and PS01). This outcome can be attributed to the different hydrologic-response times considering the location of the measurement sites. The time base of the hydrographs observed at Florida hydrometric station (Figure 2) is overall equal to 6 days and it generally do not vary with the change of the flow magnitude. Ríos [48] found that, on average, the renewal time of the Paso Severino reservoir ranges between 2 to 8 weeks. He also observed that during storm events, the renewal time can be a few days long, while it can last several months during dry periods. 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 reservoir, where the water body is lacustrine, associated with a lentic ecosystem. The validation of the imputation process was outstanding, showing overall "very good" results in terms of the PBIAS and KGE ratings.

A box-plot representation of the model performance (NSE, PBIAS, and KGE) is represented in Figure 4. More than 76% of the imputed data is characterized by NSE > 0.45 (it is at least "satisfactory"), and more than 92% of the imputed data has a positive NSE, meaning that for almost all the imputations, our methodology is better than the mean function used as an imputer. The validation results were notable. Considering PBIAS ratings, more than 96% of the imputed data can be considered at least "satisfactory" and more than 88% "very good." In terms of KGE ratings, all the imputations are considered "good."

Overall, IDW resulted in being the winning model more times compared to the others (17 times), followed by RFR (8 times), HR (6 times), SVR (5 times), and RR (4 times). This can be justified from the fact that IDW is the only model that takes into account not only the temporal information, as well as the other implemented models, but also the spatial one, by considering the neighboring stations for supporting the imputation process. The other machine-learning models implemented are more or less chosen the same number of times (same order of magnitude).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 May 2021 doi:10.20944/preprints202105.0105.v1

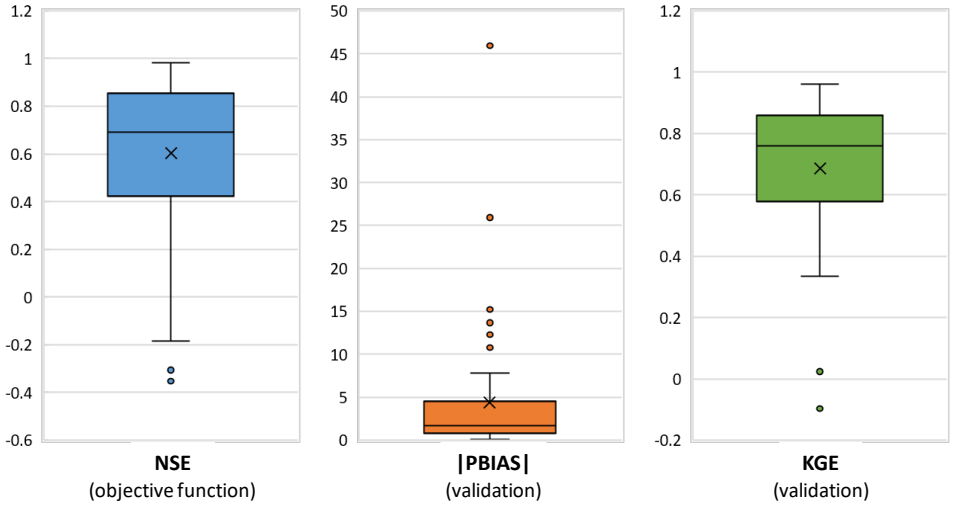

Figure 4. Box plots of model imputation performance (NSE, PBIAS, and KGE).

5. Conclusions

In this study, we tackled the challenge of imputing a multivariate water-quality dataset characterized by a high percentage of missing data (between 50% and 70%). In particular, the variables T_w, *EC*, *pH*, *DO*, *TN*, NO_2^-, NO_3^-, and *Turb* of six monitoring stations located along the Santa Lucía Chico river (Uruguay) were considered for this study. Adopting a multi-model approach was crucial since the best model for imputing any kind of water-quality variable does not exist. The statistical and machine-learning models implemented were Inverse distance weighting (IDW), Random forest regressor (RFR), ridge (RR), Bayesian ridge (BR), Adaboost (AB), Huber regressor (HR), Support vector regressor (SVR), and k-nearest neighbors regressor (KNNR).

The imputation outcomes were overall adequate. More than 76% of the imputed data can be considered "satisfactory" (NSE > 0.45). This vas validated by calculating PBIAS (> 96% of the imputed data is "satisfactory") and KGE (all the imputations are considered "good"). It is interesting to notice that the performance is always remarkable at the three monitoring stations located in the Paso Severino reservoir, while they may be "unsatisfactory" at some monitoring stations located along the Santa Lucía Chico river (upstream the reservoir). This can be attributed to the different hydrologic-response times considering the location of the measurement sites. SLC01 and SLC02 are located several kilometers upstream of the reservoir where the water body is fluvial. While PS02, PS03, and PS04 are located within the reservoir where the water body is lacustrine. Among the implemented models, IDW was chosen as the best model 17 times. We believe that the reason behind is represented by the fact that this is the only model that takes into account the temporal and spatial variability that characterizes the variable under study.

The results obtained in this study are expected to support water managers and researchers in well-exploiting the existing water-quality data to improve modeling and to generate effective pollution-control strategies.

Based on the current promising results, a further step to be considered in future works to improve the current methodology is implementing physical knowledge that takes into account the spatial information of the available water-quality data. In such a way, the current machine-learning approach will move towards a hybrid method where the data-driven techniques will be trained with physically based concepts.

Author Contributions: Conceptualization, A.G. and A.C.; methodology, R.R. and M.P.; software, R.R. and M.P.; formal analysis, R.R. and M.P.; data curation, R.R., M.P. and L.E.; writing—original draft preparation, A.G.; writing—review and editing, A.C., L.E., C.C., R.R., M.P., and M.F.; supervision, A.G., A.C., L.E., C.C., and M.F.; project administration, A.G.; funding acquisition, A.G. and A.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by ANII, grant number FSDA_1_2018_1_153967.

Data Availability Statement: The original water-quality dataset was freely downloaded from https://www.dinama.gub.uy/oan/datos-abiertos/calidad-agua/. The imputed water-quality dataset can be found in https://doi.org/10.5281/zenodo.4731169.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Whitehead, P.; Dolk, M.; Peters, R.; Leckie, H. Water Quality Modelling, Monitoring, and Management. In Water Science, Policy, and Management, eds S.J. Dadson, D.E. Garrick, E.C. Penning-Rowsell, J.W. Hall, R. Hope and J. Hughes, 2019.
2. Gorgoglione, A.; Castro, A.; Chreties, C.; Etcheverry, L. Overcoming Data Scarcity in Earth Science. *Data* **2020**, *5*, 5.
3. Teegavarapu, R.S.V.; Aly, A.; Pathak, C.S.; Ahlquist, J.; Fuelberg, H.; Hood, J. Infilling missing precipitation records using variants of spatial interpolation and data-driven methods: use of optimal weighting parameters and nearest neighbour-based corrections. *Int. J. Climatol.* **2018**, *38*, 776-793.
4. Mital, U.; Dwivedi, D.; Brown, J.B.; Faybishenko, B.; Painter, S.L.; Steefel, C.I. Sequential imputation of missing spatio-temporal precipitation data using random forests. *Frontiers in Water* **2020**, *2*, 20.
5. Aguilera, H.; Guardiola-Albert, C.; Serrano-Hidalgo, C. Estimating extremely large amounts of missing precipitation data. *J. of Hydroinformatics* **2020**, *22(3)*, 578–592.
6. Buhi, E. Out of sight, not out of mind: Strategies for handling missing data. *American Journal of Health Behavior* **2008**, *32(1)*.
7. Ratolojanahary, R.; Ngouna, R.H.; Medjaher, K.; Junca-Bourié, J.; Dauriac, F.; Sebilo, M. Model selection to improve multiple imputation for handling high rate missingness in a water quality dataset. *Expert Systems with Applications* **2019**, *131*, 299–307.
8. Lo Presti, R.; Barca, E.; Passarella, G. A methodology for treating missing data applied to daily rainfall data in the Candelaro River Basin (Italy). *Environmental Monitoring and Assessment* **2010**, *160*, 1-22.
9. Chen, F.W.; Liu, CW. Estimation of the spatial rainfall distribution using inverse distance weighting (IDW) in the middle of Taiwan. *Paddy Water Environ* **2012**, *10*, 209–222.
10. Barrios, A.; Trincado, G.; Garreaud, R. Alternative approaches for estimating missing climate data: application to monthly precipitation records in South-Central Chile. *For. Ecosyst.* **2018**, *5*, 28.
11. Gong, G.; Mattevada, S.; O'Bryant, S.E. Comparison of the accuracy of kriging and IDW interpolations in estimating groundwater arsenic concentrations in Texas. *Environmental Research* **2014**, *130*, 59-69.
12. Sattari, M.-T.; Rezazadeh-Joudi, A.; Kusiak, A. Assessment of different methods for estimation of missing data in precipitation studies. *Hydrology Research* **2017**, *48(4)*, 1032–1044.
13. Oriani, F.; Borghi, A.; Straubhaar, J.; Mariethoz, G.; Renard, P. Missing data simulation inside flow rate time series using multiple-point statistics. *Environmental Modelling & Software* **2016**, *86*, 264-276.
14. Tabari, H.; Talaee, P.H. Recontrsuction of river water quality missing data using artificial neural networks. *Water Quality Research Journal of Canada* **2015**, 50.4.
15. Srebotnjak, T.; Carr, G.; de Sherbinin, A.; Rickwood, C. A global Water Quality Index and hot-deck imputation of missing data. *Ecological Indicators* **2012**, *17*, 108–119.
16. Ratolojanahary, R.; Ngouna, R.H.; Medjaher, K.; Junca-Bourié, J.; Dauriac, F.; Sebilo, M. Model selection to improve multiple imputation for handling high rate missingness in a water quality dataset. *Expert Systems With Applications* **2019**, *131*, 299-307.
17. OAN – Observatorio Ambiental Nacional, https://www.dinama.gub.uy/oan/geoportal/, last accessed 2021/01/11.
18. Goyenola, G.; Meerhoff, M.; Teixeira-de Mello, F.; González-Bergonzoni, I.; Graeber, D.; Fosalba, C.; Vidal, N.; Mazzeo, N.; Ovesen, N.B.; Jeppesen, E.; et al. Phosphorus dynamics in lowland streams as a response to climatic, hydrological and agricultural land use gradients. *Hydrol. Earth Syst. Sci. Discuss.* **2015**, *12*, 3349–3390.
19. Aubriot, L.; Delbene, L.; Haakonson, S.; Somma, A.; Hirsch, F.; Bonilla, S. Evolución de la eutrofización en el Río Santa Lucía: Influencia de la intensificación productiva y perspectivas. *Innotec* **2017**, *14*, 7–17.
20. Gorgoglione, A.; Gregorio, J.; Ríos, A.; Alonso, J.; Chreties, C.; Fossati, M. Influence of land use/land cover on surface-water quality of Santa Lucía river, Uruguay. *Sustainability* **2020**, *12*, 4692.
21. Gorgoglione, A.; Alonso, J.; Chreties, C.; Fossati, M. Assessing temporal and spatial patterns of surface-water quality with a multivariate approach: a case study in Uruguay. *IOP Conference Series: Earth and Environmental Science* **2020**, *612*, 012002.
22. Wolpert, D.; Macready, W. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation* **1997**, *1*, 67-82.
23. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
24. Bartier, P.M.; Keller, C.P. Multivariate interpolation to incorporate thematic surface data using inverse distance weighting (IDW). *Computers & Geosciences* **1996**, *22(7)*, 195-799.
25. Tang, F.; Ishwaran, H. Random Forest Missing Data Algorithms. *Stat Anal Data Min.* **2017**, *10(6)*, 363–377.
26. Farebrother, R.W. Further results on the mean square error of ridge regression. *Journal of the Royal Statistical Society*, Series B (Methodological) **1976**, *38*, 248-250.
27. Tipping, M.E. Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research* **2001**, *1*.
28. Drucker, H. Improving Regressors using Boosting Techniques. In Proceedings of the Fourteenth International Conference on Machine Learning (ICML '97). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1997, 107–115.
29. Art, O. A robust hybrid of lasso and ridge regression. *Contemp. Math.* **2007**, *443*, 10.1090/conm/443/08555.
30. Smola, A.J., Schölkopf, B. A tutorial on support vector regression. *Statistics and Computing* **2004**, *14*, 199–222.

31. Dang, X., Peng, H., Wang X., Zhang, H. Theil-Sen Estimators in a Multiple Linear Regression Model, 2009.
32. Mucherino A., Papajorgji P.J., Pardalos P.M. k-Nearest Neighbor Classification. In: Data Mining in Agriculture. Springer Optimization and Its Applications, 34. Springer, New York, NY, 2009.
33. Narbondo, S.; Gorgoglione, A.; Crisci, M.; Chreties, C. Enhancing physical similarity approach to predict runoff in ungauged watersheds in sub-tropical regions. *Water* **2020**, *12*, 528.
34. Chen, H.; Luo, Y.; Potter, C.; Moran, P.J.; Grieneisen, M.L.; Zhang, M. Modeling pesticide diuron loading from the San Joaquin watershed into the Sacramento-San Joaquin Delta using SWAT. *Water Res.* **2017**, *121*, 374–385.
35. Gorgoglione, A.; Bombardelli, F.A.; Pitton, B.J.L.; Oki, L.R.; Haver, D.L.; Young, T.M. Role of Sediments in Insecticide Runoff from Urban Surfaces: Analysis and Modeling. *Int. J. Environ. Res. Public Health* **2018**, *15*, 1464.
36. Rogelis, M.C.; Werner, M.; Obregón, N. and Wright, N. Hydrological model assessment for flood early warning in a tropical high mountain basin. *Hydrol. Earth Syst. Sci. Discuss*. **2016**, 1–36.
37. Andersson, J.C.M.; Arheimer, B.; Traoré, F.; Gustafsson, D. and Ali, A. Process refinements improve a hydrological model concept applied to the Niger River basin. *Hydrol. Process.* **2017**, *31(25)*, 4540–4554.
38. Knoben, W.J.M.; Woods, R. A.; Freer, J.E. A Quantitative Hydrological Climate Classification Evaluated with Independent Streamflow Data. *Water Resour. Res.* **2018**, *54(7)*, 5088–5109.
39. Knoben, W.J.M.; Freer, J. E.; Woods, R.A. Technical note: Inherent benchmark or not? Comparing Nash-Sutcliffe and Kling-Gupta efficiency scores. *Hydrol. Earth Syst. Sci.* **2019**, *23*, 4323–4331.
40. Moriasi, D.N.; Arnold, J.G.; Van Liew, M.W.; Bingner, R.L.; Harmel, R.D.; Veith, T.L. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Soil Water Div. Asabe* **2007**, 50, 885–900.
41. Hayashi, M. Temperature-Electrical Conductivity Relation of Water for Environmental Monitoring and Geophysical Data Inversion. *Environ. Monit. Assess.* **2004**, *96*, 119–128.
42. Beretta-Blanco, A.; Carrasco-Letelier, L. Relevant factors in the eutrophication of the Uruguay River and the Río Negro. *Science of the Total Environment* **2021**, *761*, 143299.
43. Bakhtiar Jemily, N.H.; Ahmad Sa'ad, F.N.; Mat Amin, A.R.; Othman, M.F.; Mohd Yusoff, M.Z. Relationship Between Electrical Conductivity and Total Dissolved Solids as Water Quality Parameter in Teluk Lipat by Using Regression Analysis. In: Abu Bakar M., Mohamad Sidik M., Öchsner A. (eds) Progress in Engineering Technology. Advanced Structured Materials 2019, 119. Springer, Cham.
44. Lintern, A.; Wbb, J.A.; Ryu, D.; Liu, S.; Bende-Michl, U.; Waters, D.; Leahy, P.; Wilson, P.; Western, A.W. Key factors influencing differences in stream water quality across space. *WIREs Water* **2018**.
45. *pandas_profiling* library. Available online: https://github.com/pandas-profiling/, last accessed 2020/12/29.
46. Reisinger, A.J.; Groffman, P.M.; Rosi-Marshall, E.J. Nitrogen-cycling process rates across urban ecosystems. *FEMS Microbiol. Ecol.* **2016**, *92*, 198.
47. Refaeilzadeh, P.; Tang, L.; Liu, H. Cross-Validation. In: LIU L., ÖZSU M.T. (eds) Encyclopedia of Database Systems. Springer, Boston, MA, 2009.
48. Río, A. Implementación de un modelo hidrodinámico tridimensional en el embalse de Paso Severino. Aportes para la modelación de calidad de agua. Master Thesis. Graduate Program of Applied Fluid Mechanics, Universidad de la República, Uruguay, 2019. Available online: https://www.colibri.udelar.edu.uy/jspui/handle/20.500.12008/21553, last accessed 2021/04/29.