

Article

A Data-Centric Approach to Water Quality Prediction: Sample Size, Augmentation, and Model Performance with a Focus on Ammonium in a Tropical Wetland

Doris Mejia Avila ^{*}, Viviana Soto Barrera  and Franklin Torres Bejarano 

Environmental Engineering Department, University of Córdoba, Carrera 6 # 76-103, Monteria 230002, Colombia; vcsoto@correo.unicordoba.edu.co (V.S.B.); franklintorres@correo.unicordoba.edu.co (F.T.B.)

* Correspondence: dmejia@correo.unicordoba.edu.co; Tel.: +57-(4)-7860920

Abstract

Framed within data-centric artificial intelligence, this study integrates statistics, geotechnologies and AI to improve water quality prediction. The primary objective was to identify the minimum sample size required to train robust and accurate machine learning models. Based on 30 sampling points in a tropical wetland in northern Colombia, ammonium concentration was selected as the target variable, and total dissolved solids, suspended solids, phosphate, dissolved oxygen, nitrate and chemical oxygen demand were chosen as predictors. Because 30 observations are insufficient to train robust models, data augmentation was performed using ordinary kriging (OK) and empirical Bayesian kriging (EBK). From the kriging-interpolated surfaces, 1000 synthetic points (randomly and spatially distributed while preserving the estimated spatial structure) were sampled; from this expanded dataset, subsamples of varying sizes were drawn to train six algorithms: multiple linear regression (MLR), random forest (RF), k-nearest neighbours (k-NN), gradient boosting machines (GBM), multilayer perceptron (MLP) and radial basis function neural network (RBF-NN). The RF, k-NN, MLP, RBF-NN and GBM models trained on the interpolated data exhibited excellent performance: in the testing phase, they achieved adjusted coefficients of determination > 0.95 and symmetric mean absolute percentage errors (SMAPEs) $< 10\%$, and the resulting predictive surfaces showed comparable performance under external validation. According to the criteria of stability, goodness of fit, and external validation, the optimal minimum sample size for most algorithms was 104 observations. These results represent a significant advance in mitigating data scarcity in water quality modelling. The identification of effective data augmentation methods and the determination of appropriate sample sizes, as demonstrated here, support the robust application of AI techniques in water quality prediction. The proposed strategy is transferable to other quantitative, spatially continuous environmental variables and thus contributes to the development of the emerging subdiscipline of geospatial artificial intelligence (GeoAI).



Academic Editor: Pankaj Kumar

Received: 1 February 2026

Revised: 3 March 2026

Accepted: 5 March 2026

Published: 28 April 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: data augmentation; water quality; data-centric artificial intelligence; kriging; prediction surface

1. Introduction

Water is an essential resource for life on the planet [1]; consequently, monitoring and modelling its quality is a global challenge. Water quality monitoring encompasses numerous hydrological, physical, chemical, and biological parameters, some measured in

situ and others through laboratory analysis [2]. Water quality models have become effective tools for controlling, preventing, and remediating water resource pollution; however, their usefulness depends on accurately capturing the underlying processes; otherwise, there is a risk that management decisions will be based on unreliable information [3].

One of the primary limitations in modelling water quality is the lack of data. Refs. [4–6] indicate that traditional water quality assessment relies on manual sampling and laboratory analysis. These methods can be slow, costly, and logistically complex, particularly in remote areas. Some techniques, such as real-time in situ sensor networks (IoT—Internet of Things) and remote sensing platforms that are integrated with artificial intelligence, provide more comprehensive data. However, when this technology is unavailable or when the aim is to predict the water quality of a specific water system at a given time, in situ monitoring remains essential for modelling. Ref. [7] emphasises that the scarcity of data in water quality assessments is a significant problem that must be addressed.

In current AI-based modelling, the pattern recognition capacity of machine learning is proportional to dataset size: the smaller the dataset, the less powerful and accurate the algorithms [8]. Therefore, the use of small datasets is a limitation. Ref. [9] notes that ML research has focused mainly on developing model types and architectures and on defining hyperparameters to improve performance; however, in machine learning, the algorithm is not the principal factor determining the accuracy of the final classification [10].

Ref. [11] defined data-centric artificial intelligence (AI) as an emerging field focused on data engineering to improve AI systems by increasing data quality and quantity. This approach emphasises the development and application of methods, tools, and best practices to systematically design datasets that optimise the performance of AI-based systems [12]. The data-centric AI approach comprises three objectives [11]: (1) training data development, which encompasses data collection, labelling, preparation, cleaning, reduction, and augmentation; (2) inference data development, which is geared towards model deployment; and (3) data maintenance, which ensures data quality and reliability in dynamic environments.

Data augmentation addresses two central research challenges: it increases data volume from limited datasets and reduces overfitting [13]. Methodological advances have focused primarily on object detection and medical image analysis [14,15].

In enhancing data for water quality assessment, recent studies have employed various methods. Ref. [7] assessed the water quality of the Selangor and Skudai rivers in Malaysia using four techniques to generate synthetic data from actual observations: the conditional tabular generative adversarial network (CTGAN), the tabular variational autoencoder (TVAE), the Gaussian copula (GC), and the copula generative adversarial network (CopulaGAN). The authors found that the estimates that were derived from the augmented datasets were more accurate than those obtained from the smaller, real datasets.

Ref. [16] asserts that spatial interpolation methods can serve as effective data augmentation techniques because they accurately capture the stochasticity, dependence, and spatial variability inherent in environmental phenomena. The authors assessed the efficiency of the OK interpolation method as a means of enhancing the accuracy of models that predict the total dissolved solids (TDS) water quality parameter. They found a statistically significant difference between models trained with field samples and those trained with samples derived from data augmentation. Consequently, they concluded that the ordinary kriging method is an effective technique for improving a model's quality.

Ref. [17] employed two interpolation methods—Gaussian processes (GP) with different kernels and kriging using various variograms—to augment the number of observations in an agricultural database aimed at predicting the coverage of the weed *Commelina benghalensis* L. (COMBE) in sugarcane plots. Ref. [18] generated pseudo-labels from field measurements using conventional kriging and subsequently combined these with unla-

belled satellite images. They then integrated the enriched images with an existing training set containing actual field measurements. Using this enhanced dataset, they trained a convolutional neural network model—the CNN-RF model—and compared its performance against a model trained without pseudo-labels.

Having sufficient data is essential; however, determining the appropriate sample size is crucial for the correct and objective evaluation of AI algorithms [19]. Oversampling can lead to unnecessary expenditure of time and resources [20]. The influence of sample size on model performance remains an area of ongoing research that requires further investigation across most fields [21].

Currently, research on sample size requirements for AI applications is primarily focused on clinical contexts. For instance, ref. [20] proposed a procedure for estimating sample size in the development of AI models, which comprises the following steps: (a) define the error threshold or clinically acceptable accuracy; (b) construct multiple subsets with smaller sample sizes using random resampling with several repetitions; (c) train the models on each subset; and (d) estimate the optimal data size using a curve that relates model performance to different sample sizes.

The scarcity of data for assessing water quality is a persistent limitation in tropical regions, where the absence of real-time acquisition technologies and budgetary constraints hinder the collection of sufficient in situ data series. This lack of data restricts the application of AI in water quality modelling and impedes the development of robust models that can provide reliable information for water resource management and conservation.

This study, situated within a data-centric AI framework, addresses the scarcity of in situ observations and the determination of the optimal sample size for modelling water quality. The primary objective was to identify the minimum sample size required to train robust machine learning models capable of accurately predicting water quality. To increase data availability, OK and EBK were applied to a sample of in situ water quality measurements.

The dataset derives from sampling conducted in situ on 27 September 2024, a date that, in the Bajo Sinú region (Lower Sinú), coincides with the rainy season, when the wetland complex attains its maximum storage capacity. Seventeen water quality parameters were recorded at 30 locations distributed throughout the Guartínaja wetland.

From this sample the following procedure was applied: (1) ammonium (NH_3) was defined as the target parameter; (2) dimensionality reduction techniques were used to identify the best predictors; (3) data augmentation was performed using OK and EBK; (4) for each kriging surface a sample of 1000 points was extracted and subsamples of varying sizes were generated; (5) five AI regression algorithms were implemented in Python: multiple linear regression (MLR), random forest (RF), k-nearest neighbours (k-NN), gradient boosting machines (GBM), multilayer perceptron (MLP) regressor, and radial basis function neural network (RBF-NN); (6) each algorithm was trained on the augmented subsamples; and (7) sample sizes were compared by evaluating model stability, explanatory power, and external validation of prediction surfaces against in situ data.

Geostatistical methods—OK and EBK—proved suitable as data augmentation strategies. The RF, k-NN, MLP, RBF-NN and GBM models that were trained on the interpolated data exhibited high explanatory power, with adjusted coefficients of determination (a_R^2) exceeding 0.95 and symmetric mean absolute percentage errors (SMAPEs) below 10% in both testing and on the derived prediction surfaces.

In evaluating different sample sizes for training and validation, the stabilisation point for the five models was reached with 200 observations. Under the criteria of stability, goodness of fit, and external evaluation, the optimal minimum sample size for most algorithms was 104 observations. There were two exceptions: (a) the MLR algorithm, whose SMAPE values in both the testing phase and external evaluation exceeded 10%

and (b) the RBF-NN algorithm, which met all three conditions with a sample size of 78 observations, indicating greater explanatory power with a smaller sample.

Framed within data-centric AI, this study proposes an integrated strategy combining statistical tools, geotechnologies and AI to produce more robust models with improved predictive accuracy for describing the state of aquatic ecosystems. Identifying effective data augmentation methods and appropriate sample sizes, as demonstrated here, will facilitate a broader adoption of AI for water quality prediction. Future research should extend this approach to other water quality indicators. The methodology is also transferable to other quantitative, spatially continuous environmental variables, thereby contributing to the development of the emerging geospatial artificial intelligence (GeoAI) subdiscipline.

2. Materials and Methods

2.1. Study Area

In situ sampling was conducted on 27 September 2024 in the Guartínaja wetland, located within the Bajo Sinú Wetland Complex (BSWC) in the Caribbean region of northern Colombia. This complex is the lowest-lying fluvial–lacustrine unit in the Sinú River basin [22] and comprises swamps, estuaries, and freshwater ponds, both permanent and temporary. The BSWC naturally buffers and regulates rainfall and excess water during Sinú River floods [23]. The study area has a mean temperature of 27.5 °C, mean precipitation of 1253.3 mm, and mean evaporation of 1007.1 mm [24]. Figure 1 shows the geographical location of the Guartínaja wetland.

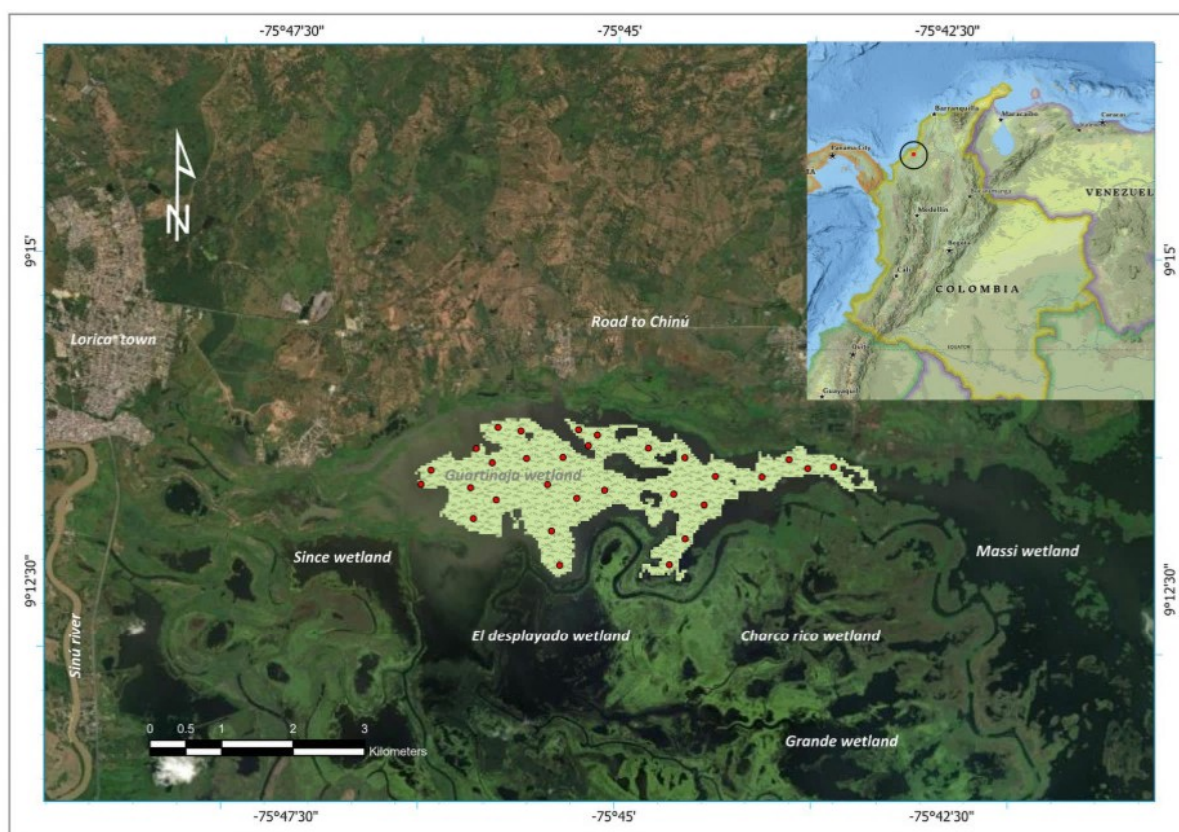


Figure 1. Geographic location of the Guartínaja wetland within the BSWC.

2.2. Dataset

During the sampling campaign on 27 September 2024, water quality data for 17 parameters were collected at 30 locations within the wetland (Figure 1). This date,

in the Bajo Sinú region, coincides with the rainy season, when the wetland complex attains its maximum storage capacity.

The locations of the field observations are detailed in Appendix A. The parameters measured included: turbidity; temperature; total dissolved solids; total suspended solids; suspended solids; salinity; total phosphorus; dissolved phosphates; organic nitrogen; nitrogen dioxide; ammonia; pH; dissolved oxygen; nitrate; chemical oxygen demand; biochemical oxygen demand; and electrical conductivity. Table 1 presents the parameters along with their respective units of measurement and methods of collection.

Table 1. Water quality parameters collected in the Guartínaja wetland during the September 2024 sampling campaign.

Parameter Name	Initials	Unit of Measurement	Method of Obtaining
Hydrogen ion potential	pH		Multiparameter probe for water quality
Dissolved oxygen	DO	mg/L	
Electrical conductivity	EC	mU/cm	
Total dissolved solids	TDS	mg/L	
Turbidity	TURB	NTU	
Biological availability of oxygen	BAO	mg/L	Laboratory
Chemical availability of oxygen	DOC	mg/L	
Nitrate ion	NO ₃	mg/L	
Nitrite ion	NO ₂	mg/L	
Ammonia	NH ₃	mg/L	
Total phosphorus	PT	mg/L	
Phosphate	PO ₄	mg/L	
Organic nitrogen	NORG	mg/L	
Total suspended solids	SST	mg/L	
Salts	SAL	UPS	Multiparameter probe for water quality
Suspended solids	SS	mg/L	Laboratory
Temperature	T	°C	Multiparameter probe for water quality

For in situ measurements, a HI-9829 multiparameter probe (HANNA, Leighton Buzzard, UK) was used, following the technical specifications of the Colombian Institute of Hydrology, Meteorology and Environmental Studies (IDEAM) for field data collection. In accordance with these specifications, the equipment was calibrated and checked prior to each measurement following the manufacturer's instructions. During each reading, the probe was rinsed with the sample, and the multiparameter electrode was cleaned with distilled water after recording the data to prevent cross-contamination between measurements [25].

2.3. Dimensionality Reduction in the Sample

The parameter selected as the target variable was (NH₃). NH₃ is a by-product of the protein metabolism excreted by fish and of the bacterial degradation of organic matter (food debris, faeces, dead plankton, sewage, among others). In its unionised form (NH₃), it is highly toxic [26] and is found predominantly in surface waters.

To identify suitable predictors for NH₃ modelling, the following procedure was applied. A Pearson correlation analysis was performed between NH₃ and the candidate predictors. Pearson's correlation coefficient [27] measures the strength and direction

of the linear relationship between two variables, with values ranging from -1 (perfect negative correlation) to $+1$ (perfect positive correlation). Predictors with an absolute correlation $|p| \geq 0.50$ were retained. The equation used to calculate Pearson's correlation is presented below.

$$p(a, b) = \frac{\sigma_a^2 \times \sigma_b^2}{\sigma_a \times \sigma_b} \quad (1)$$

In this equation, $p(a, b)$ represents the Pearson's correlation, σ_a^2 represents the standard deviation of parameter a and σ_b^2 represents the standard deviation of parameter b .

For parameters correlated with NH_3 , the variance inflation factor (VIF) test [28] was employed to assess the degree of multicollinearity between the predictive parameters and the target parameter. The equation used to calculate the VIF is presented below.

$$VIF = \frac{1}{1 - R_x^2} \quad (2)$$

where $(1 - R_x^2)$ is the coefficient of determination of the linear regression between two predictive parameters. A VIF of 1 indicates that there is no multicollinearity between the two compared variables; a VIF between 1 and 5 suggests moderate multicollinearity, which is generally acceptable; and a VIF greater than 5 indicates high multicollinearity [29]. From the pairs of predictive parameters that exhibited a VIF value greater than 5, the parameter with the lowest correlation to the target parameter (NH_3) was discarded.

2.4. Data Augmentation

2.4.1. Analysis of the Suitability of Applying Geostatistical Methods

Spatial interpolation requires the sample data to be interpolated and for it to exhibit a Gaussian distribution and spatial autocorrelation [30]. For each parameter, a normality analysis was conducted based on the histogram [31], defining a normalisation function for those parameters that did not conform to a normal distribution. To determine the degree of spatial correlation for each in situ dataset, the Global Moran's I test was applied. This test utilises the locations and values of the analysis parameter to ascertain whether the pattern of a dataset is dispersed or clustered [32]. The Global Moran's I index (I_{moran}) is calculated using the following equation [33]:

$$I_{\text{moran}} = \frac{n \sum_{x=1}^n \sum_{y=1}^n w_{x,y} z_x z_y}{So \sum_{x=1}^n z_x^2} \quad (3)$$

where z_x represents the deviation of an attribute from its mean ($x_i - \bar{X}$). For feature X , $w_{x,y}$ is the spatial weight between features x and y , n is the total number of features, and So is the sum of all the weights of the spatial features. The range of Moran's I index is between $+1$ and -1 . A value of $+1$ suggests a clustering pattern of the data, as 0 indicates no correlation and -1 signifies a dispersed pattern [33].

Spatial autocorrelation was reaffirmed through a spatial randomness hypothesis test, where the null hypothesis posits that the data exhibit a random spatial distribution. The test was assessed based on the standard deviation (z-score) and the probability (p -value) to determine the reliability of the data autocorrelation [32]. Z-score values greater than -1.65 and p -values less than 0.05 led to the rejection of the null hypothesis [34].

2.4.2. Application of Geostatistical Methods for Data Augmentation

Two geostatistical methods were employed: OK and empirical EBK. For each parameter, both methods were applied, and the kriging surface exhibiting the best statistical performance was selected.

OK is a stochastic method that assumes the distance or direction between known points and reflects the spatial correlation used to explain surface variation [35].

In OK, the weights are determined not only by the distance between the measured points and the prediction location but also by the overall spatial arrangement of those points [36]. The calculation of unknown values is based on the calculation of the semi-variogram or the experimental model (Equation (2)), which is subsequently adjusted to a theoretical model from which the unknown values are calculated [37].

$$\gamma(h) = 1/2 \left(\sum_{i=1}^{n(h)} \frac{(z(x+h) - z(x))^2}{n(h)} \right) \quad (4)$$

where $\gamma(h)$ is the semivariance, $z(x+h) - z(x)$ is the variance of the dataset at the same distance and $n(h)$ is the number of points at distance h .

In the empirical EBK method, the error introduced when estimating the semivariogram model is considered, for which several semivariogram models are utilised [37]. Three criteria proposed by [38] are followed: (a) the model estimates the semivariogram from the input data; (b) a new value is predicted for each location in the dataset based on the semivariogram; and (c) a model is calculated from the simulated dataset.

Concentration values for each parameter were extracted from the kriging surfaces at 1000 locations distributed throughout the Guartínaja wetland area, excluding the locations from the original sample. This dataset constitutes the augmented sample.

2.5. Machine Learning (ML) Algorithms

Six regression algorithms, built with the Scikit-learn (<https://scikit-learn.org/stable/> (accessed on 5 May 2025)) ML library, were employed: MLR adjusted through the OLS method, RF, k-NN, GBM, MLP regressor, and RBF-NN.

2.5.1. Multiple Linear Regression (MLR)

To develop the MLR algorithm, the ordinary least squares (OLS) method was employed, which fits a regression plane to a ‘cloud’ of data that is assumed to exhibit a linear trend [39]. The Statsmodels library in Python 3.13 (https://www.statsmodels.org/dev/generated/statsmodels.regression.linear_model.OLS.html (accessed on 20 May 2025)) was utilised, allowing for the evaluation of the validity, robustness, and compliance with the assumptions of the regression model, such as normality, independence, homoscedasticity of the residuals, and multicollinearity of the predictor variables.

2.5.2. Random Forest (RF)

The RF algorithm, proposed by [40], is a general-purpose classification and regression method that operates on the simple yet effective principle of ‘divide and conquer’: a sample fraction of the data is taken, a random tree predictor is created for each subset, and then these predictors are aggregated [41]. Random decision forests easily adapt to the non-linearities present in the data and, as a result, tend to provide more accurate predictions than linear regression [42].

2.5.3. k-Nearest Neighbours (k-NN)

The k-NN algorithm is a non-parametric, instance-based supervised machine learning method utilised for classification and regression tasks [43].

2.5.4. Gradient Boosting Machines (GBM)

This algorithm constructs new base learners that are maximally correlated with the negative gradient of the loss function associated with the entire dataset [44]. GBMs have demonstrated considerable success, not only in practical applications but also in various ML and data mining challenges [45].

2.5.5. Multilayer Perceptron Regression (MLP Regressor)

An MLP is a neural network composed of layers of interconnected and interchangeable artificial neurons. The number of layers in a neural network corresponds to the number of perceptron layers [46].

2.5.6. Radial Basis Function Neural Network (RBF-NN)

In an RBF-NN, there is only one hidden layer between the input and output layers. In the hidden layer, the independent variables (provided in the input layer) are fed into a set of multivariate basis functions to form the hidden features (i.e., the outputs of the hidden layer). Subsequently, in the output layer, these features are weighted and linearly summed (i.e., RBF-NN approximation) to approximate the target dependent variables [47].

2.6. Construction of the Learning Curve

With the increased sample size of 1000 data points, the six machine learning algorithms were trained to generate learning curves, determining breakpoints at intervals of 20 data points. A three-fold cross-validation method was employed for repetition. The values of the predictive variables were normalised using the MinMaxScaler from the Scikit-learn library (<https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html> (accessed on 10 June 2025)). To ensure comparability between models, the same initialisation seed (random state) was used.

To determine the optimal hyperparameters, the GridSearchCV tool from the Scikit-learn library (https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html (accessed on 10 June 2025)) was combined with a manual adjustment. The explanatory power of the models was evaluated using the following statistics:

- (a) a_R^2 . The adjusted coefficient of determination assesses the model's capacity to explain the target variable while penalising the complexity of the model, which is associated with the number of data points and predictor variables [48]. The equation for calculating the coefficient of determination is as follows:

$$a\_R^2 = 1 - \left[\left(\frac{n-1}{n-k-1} \right) \right] \times (1 - R^2) \quad (5)$$

where

a_R^2 = Adjusted coefficient of determination

n = number of observations (sampled points)

k = number of independent variables.

- (b) Root mean square error (RMSE) [49]. This is the square root of the average of the squared differences between the predicted values and the actual observed values. RMSE penalises larger errors [50] and is calculated using the following equation (Equation (4)):

$$RMSE = \frac{1}{n} \sqrt{\sum_i^n (X_i - Y_i)^2} \quad (6)$$

X_i is the predicted value, Y_i is the actual value, and n is the number of data points.

- (c) Symmetric mean absolute percentage error (SMAPE) [51]. It was calculated using the following equation:

$$\frac{100\%}{m} \times \sum_{i=1}^m \frac{|X_i - Y_i|}{\frac{(|X_i| + |Y_i|)}{2}} \quad (7)$$

where:

X_i = measured value

Y_i = predicted value

m = number of data points

SMAPE is useful for comparing the predictive performance of regression models across two different datasets with varying scales of values. Cross-validation and test statistics were conducted on the original data scale.

2.7. Training Models with Varying Sample Sizes

Based on the learning curves, we identified the sample size at which performance plateaued (stabilisation peak) and used it to define the subsamples. The subsamples were drawn from the expanded dataset, with each one ensuring proper distribution across the entire wetland area. The training procedure and evaluation metrics for the subsamples matched those used to generate the learning curves.

2.8. Selection of the Optimal Minimum Sample Size for Generating Predictive Surfaces for NH_3

The methodology for selecting the minimum sample size was adapted following the guidelines of [20] for estimating sample size when developing an AI model. Three mandatory conditions were established:

Condition 1. *Model stability. Performance must be invariant across different data partitions. To evaluate this, the difference in a_R^2 between training and validation and between training and testing must be less than 5%.*

Condition 2. *Model fit and accuracy. Models were selected that met: (a) in the testing phase, $a\_R^2 > 0.8$ and SMAPE < 10%; and (b) at a performance superior to the arithmetic mean [52], model RMSE less than the standard deviation of the sample from which it was derived.*

Condition 3. *External Validation: For the models that satisfied Conditions 1 and 2, prediction surfaces were generated, and the predicted values were compared with in situ measurements. Models were selected based on the criterion that their SMAPE value was below 10%, calculated from the predictions and field measurements.*

To construct the prediction surfaces, the input parameters (predictor variables) were the pixel values of the kriging surfaces, excluding the 1000 data points from the augmented sample to prevent dependence on the training dataset.

The optimal minimum sample size for generating predictive surfaces for NH_3 was identified as the size that trained the greatest number of algorithms meeting the three proposed conditions.

2.9. Prediction Surfaces

The prediction surfaces that satisfied the three proposed conditions were selected and subsequently compared by calculating the spatial correlation matrix.

Figure 2 presents a summary diagram of the methodology.

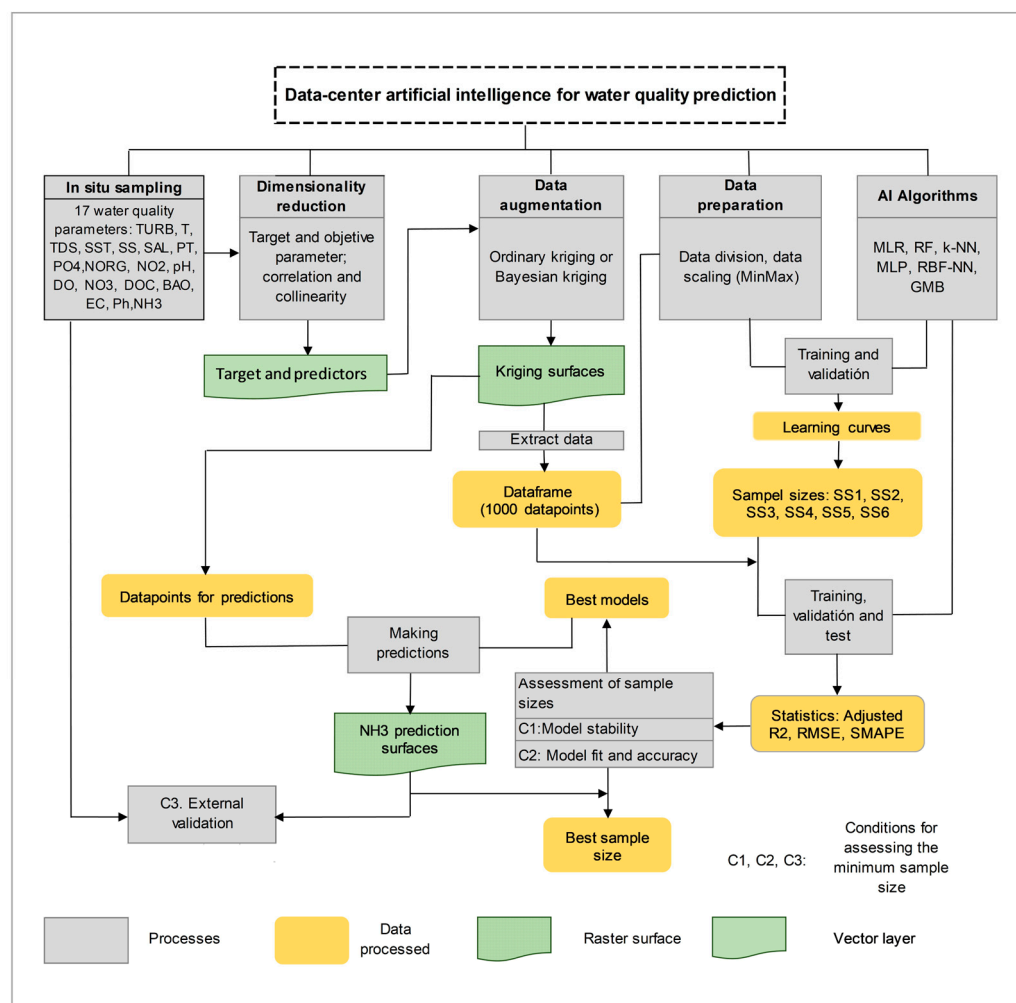


Figure 2. Summary diagram of the methodology.

3. Results

3.1. Reduction in Sample Dimensionality

Of the parameters measured in the field, nine showed a correlation greater than ± 0.5 with the NH_3 parameter: TDS, SS, SAL, PO_4 , DO, NO_3 , COD, BOD and EC. Table 2 presents the correlation matrix between NH_3 and the other parameters.

Table 2. Correlation between NH_3 and the other parameters.

Parameter	Correlation
TURB	−0.273982
T	−0.248871
TDS	0.796900
SST	−0.054644
SS	0.599073
SAL	0.70410
PT	0.482720
PO_4	0.562322
NORG	0.362493
NO_2	0.409802
pH	−0.255965
DO	−0.752283

Table 2. *Cont.*

Parameter	Correlation
NO ₃	−0.507198
COD	0.820505
BAO	0.725846
EC	0.807020

Table 3 presents the VIF matrix for the nine parameters correlated with NH₃. The SAL, EC and BOD parameters showed VIF values greater than five, indicating collinearity among them; consequently, they were discarded as predictors. The final predictors selected were TDS, SS, PO₄, DO, NO₃ and COD.

Table 3. VIF matrix for the nine parameters correlated with NH₃.

	TDS	SS	SAL	PO ₄	DO	NO ₃	COD	BAO	EC
TDS		2.6	6.2	2.2	1.3	1.4	3	2	11.1
SS	2.6		2.6	7.2	1.1	2.8	1.3	1.1	2.6
SAL	6.2	2.6		2	1.3	1.3	2	1.5	7.6
PO ₄	2.2	7.2	2		1	3.8	1.6	1.2	2.2
DO	1.3	1.1	1.3	1		1.1	1.6	1.5	1.3
NO ₃	1.4	2.8	1.3	3.8	1.1		1.3	1.1	1.4
COD	3	1.3	2	1.6	1.6	1.3		23.6	3.2
BAO	2	1.1	1.5	1.2	1.5	1.1			2
CE	111	2.6	7.6	2.2	1.3	1.4	3.2	2	

Table 4 presents the descriptive statistics of the in situ sample for the target parameter and the predictors. A total of 30 measurements were available for all parameters. The parameter with the highest variability was NH₃, with a coefficient of variation (CV) of 85.7%.

Table 4. Basic statistics of the in situ sample for the target parameter and predictors.

Statistical	NH ₃	COD	NO ₃	DO	PO ₄	SS	TDS
Number of data points	30	30	30	30	30	30	30
Null data	0	0	0	0	0	0	0
Minimum value	0.02	12.57	0.43	4.26	0.06	0.48	111.64
Maximum value	0.21	39.21	1.28	6.24	0.17	0.55	139.13
Mean	0.07	25.50	0.88	5.43	0.09	0.51	124.65
Median	0.05	25.45	0.96	5.55	0.08	0.51	124.06
Standard deviation (SD)	0.06	8.01	0.29	0.55	0.03	0.02	6.89
Coefficient of variation (CV)	85.7	31.4	33.0	10.1	33.3	3.9	5.5
Skewness	0.4	0.07	−0.17	−0.88	1.04	0.33	0.44
Kurtosis	2.58	2.14	1.46	2.76	3.55	2.1	2.76

3.2. Data Augmentation

Analysis of Sample Suitability for the Application of Geostatistical Techniques

Table 5 presents the statistics resulting from the I_{moran} test and the spatial randomness hypothesis test.

Table 5. Statistics resulting from the I_{Moran} test and the spatial randomness hypothesis test.

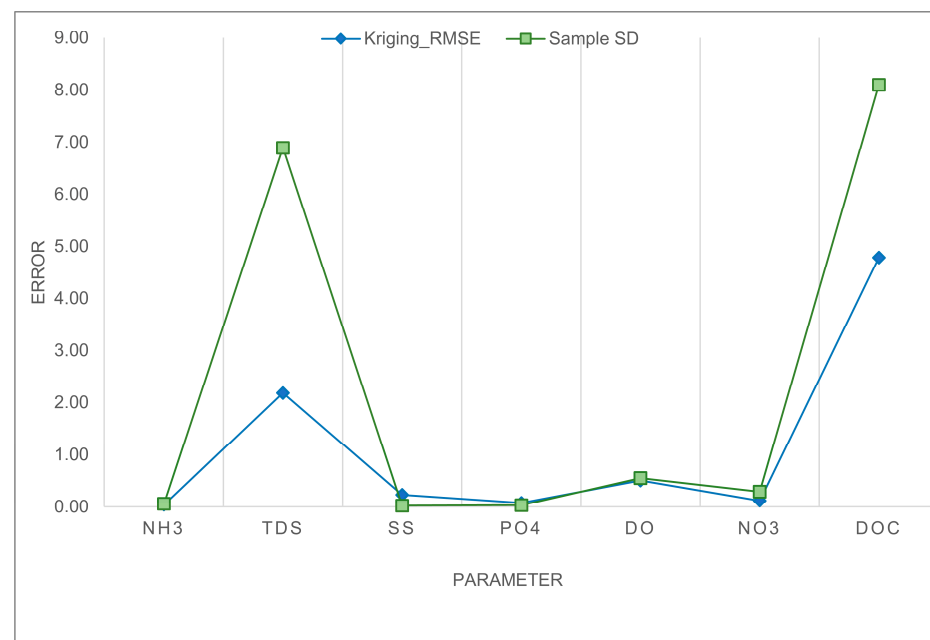
Statistical	NH ₃	COD	DO	NO ₃	PO ₄	SS	TDS
Number of data points	30	30	30	30	30	30	30
Moran Index	0.97	0.66	0.92	0.96	0.87	0.87	0.692
Z_score	5.33	3.86	4.97	5.43	4.73	5.05	4.1
p_value	0.000000	0.000112	0.000001	0.000000	0.000002	0.000000	0.000041
Spatial correlation	High	High	High	High	High	High	High
Normalisation function	Logarithmic		Square root	Logarithmic			

An I_{Moran} value greater than 0, z-scores exceeding 1, and p -values well below 0.05 indicate strong and statistically significant spatial autocorrelation in the sample. These results indicate that geostatistical methods can be applied to the in-situ samples.

3.3. Geostatistical Interpolation

Data augmentation for the parameters NH₃, COD, DO, NO₃, and SS was conducted utilising OK methodology. In contrast, EBK was employed for the parameters TDS and PO₄. Appendix B Table A1 compares the RMSE from the cross-validation of each kriging method with the standard deviation of the field sample, demonstrating that the interpolated estimates replicate the original sample's variability. The interpolation reports detailing these analyses are available in Appendix B Table A2.

Figure 3 shows a graphical comparison between the RMSE of cross-validation and the standard deviation of the in-situ data. For most parameters, the interpolation adequately reproduces the intrinsic variability of the data, given that the prediction error was lower than or very close to the standard deviation of the sample.

**Figure 3.** Comparison between the RMSE of cross-validation and the standard deviation of the in-situ data for the predictor parameters and the target parameter.

3.4. Augmented Sample

From each kriging-derived surface, 1000 locations uniformly distributed across the wetland were sampled and their estimated values extracted. The descriptive statistics of the augmented sample for each parameter are presented in Table 6.

Table 6. Descriptive statistics of the augmented sample for each parameter.

Statistical	NH ₃	COD	NO ₃	DO	PO ₄	SS	TDS
count	1000	1000	1000	1000	1000	1000	1000
min	0.02	11.08	0.42	4.25	0.06	0.48	111.04
max	0.21	40.19	1.30	6.30	0.17	0.55	139.99
range	0.20	29.11	0.88	2.04	0.11	0.08	28.95
mean	0.06	24.09	0.88	5.54	0.09	0.51	123.45
median	0.05	24.65	0.91	5.68	0.08	0.50	122.73
SD	0.05	7.90	0.27	0.50	0.03	0.02	6.37
CV	77.3	32.8	31.2	9.1	30.0	3.4	5.2

Figure 4 shows the comparison between the in-situ sample and the augmented sample using the mean (a) and CV (b). The overlap of the graphs for both datasets, for both the mean and the coefficient of variation, highlights the similarity between the two samples.

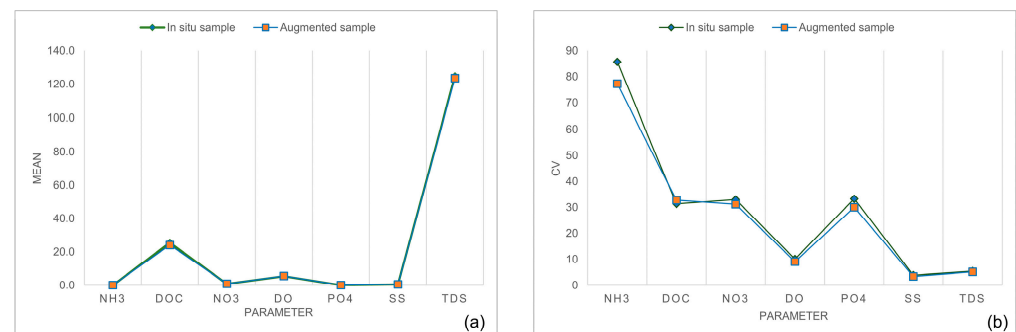


Figure 4. Comparison between the in-situ sample and the enlarged sample using the statistics of mean (a) and the coefficient of variation (b).

3.5. Configuring Sample Sizes

Figure 5 shows the learning curves for the models trained with the augmented sample. In cross-validation, four folds were used.

The learning curves of the models trained with the augmented sample illustrate that performance during the training phase (represented by the blue curve) was high and relatively stable, with the exception of the MLR model. The shaded area for the MLR model indicates a degree of uncertainty during the training phase. In contrast, during the validation phase (represented by the yellow curve), models trained with smaller sample sizes (20 and 40) exhibit a decline in performance, indicative of increased instability (shaded areas).

The six algorithms reached their stabilisation point with a sample size of around 200 observations; that is, above this threshold, improvements in performance are minimal. With sample sizes below 200, some algorithms showed instability. For example, in the MLP, the wide shading in the training curve indicates ambiguity or variation in the values of a_{R^2} , and, additionally, the validation performance is considerably lower than the training performance (0.9 and 0.7, respectively). Similarly, the MLR models with sample sizes below 100 show ambiguity in the values of a_{R^2} during the training phase.

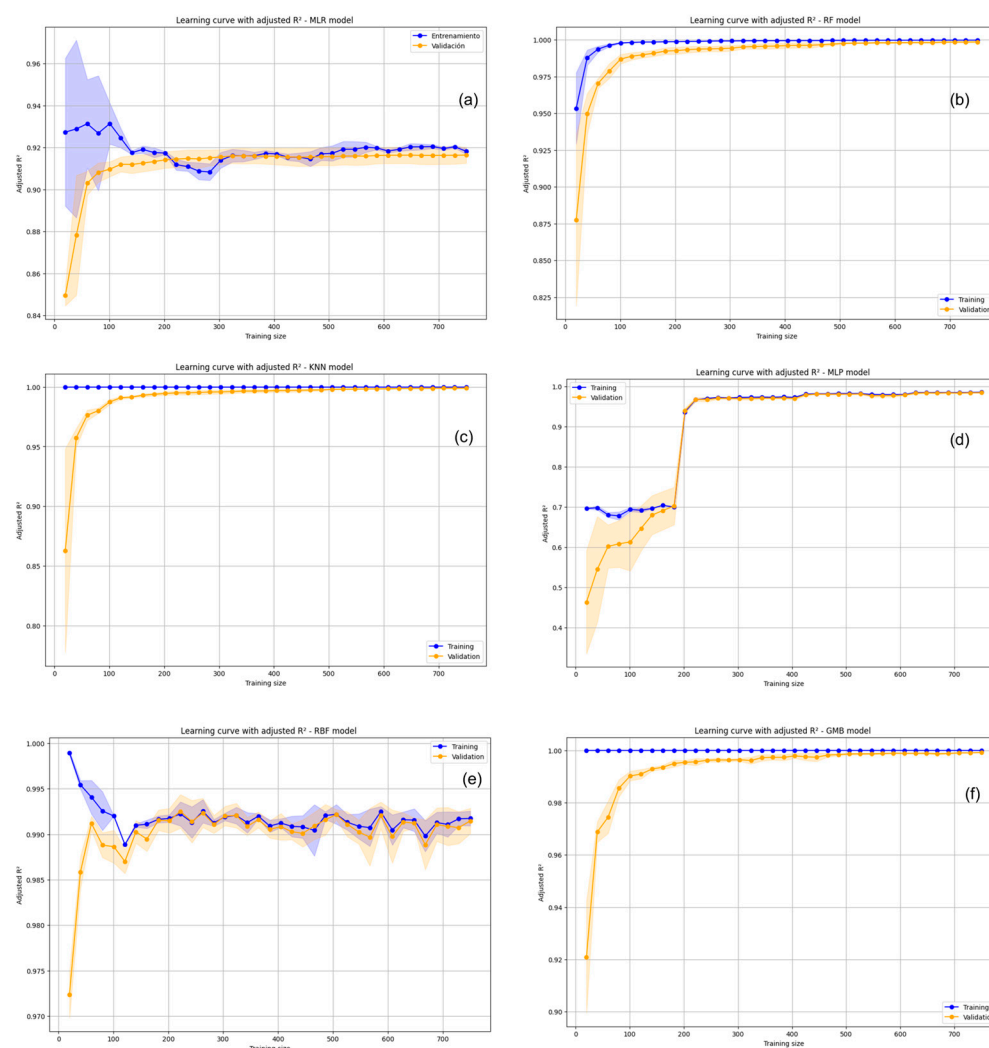


Figure 5. Learning curves for models trained with the augmented sample. Algorithms: (a) MLR, (b) RF, (c) k-NN, (d) MLP, (e) RBF-NN, and (f) GMB. The shaded areas denote the level of uncertainty associated with the model.

Table 7 lists the sample sizes used for training, cross-validation and testing, considering 200 data points as the maximum training size. For each training size, an additional 30% of data was added for the testing phase. A total of 48 models were trained: eight sample sizes for each model type, including the complete set of 1000 data points.

Table 7. Sample sizes for training, cross-validation (k-folds = 3), and testing.

Number of Data Points for Training and Validation	Testing Phase (Number of Data Points)	Sample Size (SS)
20	6	26
40	12	52
60	18	78
80	24	104
100	30	130
150	45	195
200	60	260
1000	700	300

Table 8 presents the descriptive statistics for the NH_3 parameter for each sample size.

Table 8. Objective parameter statistics for each sample size.

	NH ₃ _26	NH ₃ _52	NH ₃ _79	NH ₃ _104	NH ₃ _130	NH ₃ _195	NH ₃ _260
number of data points	26	52	79	104	130	195	260
min	0.0192	0.0165	0.0165	0.0164	0.0164	0.0164	0.0164
max	0.2138	0.2134	0.2136	0.2124	0.2099	0.2124	0.2138
range	0.1946	0.1969	0.1970	0.1960	0.1935	0.1960	0.1974
sum	1.8291	3.6717	5.0973	7.2833	8.2470	12.6261	17.3233
mean	0.0703	0.0706	0.0645	0.0700	0.0634	0.0647	0.0666
median	0.0507	0.0516	0.0488	0.0493	0.0480	0.0453	0.0468
SD	0.0571	0.0549	0.0495	0.0568	0.0477	0.0518	0.0539
CV	0.8111	0.7769	0.7664	0.8109	0.7513	0.8002	0.8091

3.6. Training ML Algorithms Using Different Sample Sizes

Table 9 shows the set of hyperparameters used to train each type of algorithm. In the case of the MLR algorithm, no hyperparameters were used. Tables A4–A8 in Appendix C provide detailed information on the hyperparameter values used in each of the 48 models. Each model is named using an ML algorithm acronym, followed by a number indicating the sample size for training, validation and testing.

Table 9. Hyperparameters used in training each type of algorithm.

Learning Algorithm	Hyperparameters of the Models
RF	max_depth. 'max_features. min_samples_leaf. min_samples_split. n_estimators.
k-NN	algorithm. leaf_size. metric. n_neighbors. p. weights
MLP	activation. alpha. hidden_layer_sizes. learning_rate. max_iter. solver
RBF-NN	num_centers. sigma
GMB	learning_rate. max_depth. max_features. min_samples_leaf. min_samples_split. n_estimators. subsample

The performance comparison in the training, validation, and testing phases, broken down by algorithm and sample size, is shown in Figure 6. None of the models trained with a sample size of 26 observations achieved an a_R^2 greater than 0.5 in the testing phase; therefore, all models with this sample size were excluded from the analysis.

A pattern of improved fit was observed with increasing sample size. All models exhibited high fit during the training phase (fit index > 0.90), with lower values in the validation and testing phases.

The models with a sample size of 52 observations exhibited the greatest variation in fit across the training, validation, and testing phases. The differences between training and testing were as follows: MLR = 0.39; RF = 0.25; k-NN = 0.17; MLP = 0.096; RBF-NN = 0.092; and GMB = 0.18. These results suggest that, at least for the MLR, RF, and k-NN models, the underlying variability of the sample was not adequately captured, or there is overfitting, given that performance is high in training but declines when predicting the target variable with unseen data.

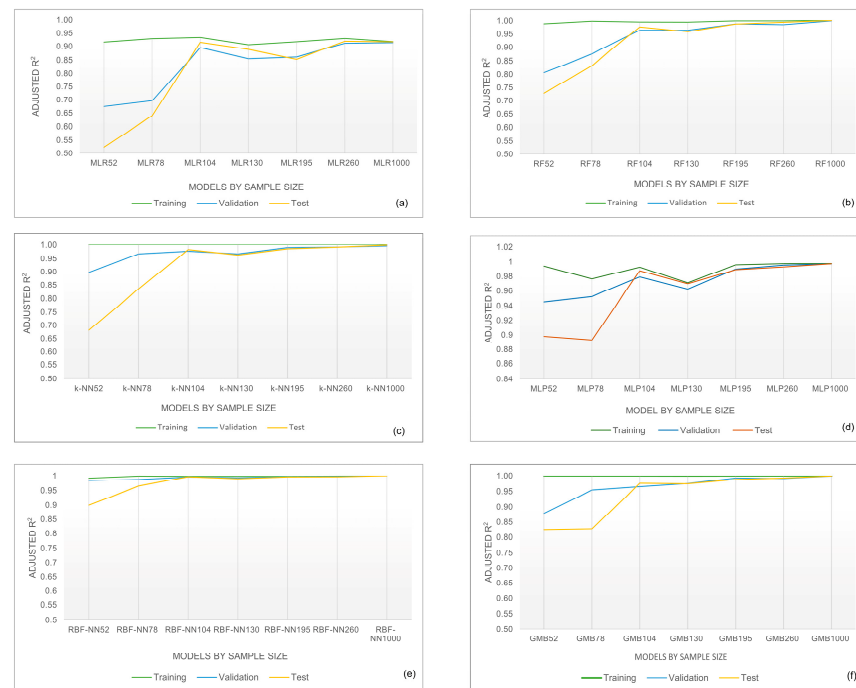


Figure 6. Comparison of performance across the training, validation, and testing phases by algorithm and sample size: (a) MLR; (b) RF; (c) k-NN; (d) MLP; (e) RBF-NN; and (f) GMB.

With a sample size of 104 observations, the separation between the three curves (training, validation, and testing) narrows, indicating stability and adequate generalisation capacity of the models. a_R^2 values ≥ 0.95 across all three phases demonstrate high predictive performance.

Figure 7 compares a_R^2 values in the testing phase across models with varying sample sizes. The RBF-NN algorithm consistently outperforms the others at all sample sizes, whereas MLR exhibits the lowest performance. The remaining algorithms show similar performance for sample sizes of 78 observations or greater.

Figure 8 compares SMAPEs in the testing phase for algorithms across different sample sizes. The MLR algorithm exhibits the largest errors and is the only one that does not show improvement as the sample size increases. RBF-NN consistently yields the lowest errors across all sizes. From a sample size of 104 observations, except for MLR, the errors of the other algorithms are less than 10%.

3.7. Selecting the Optimal Sample Size

3.7.1. Analysing Model Stability (Condition 1)

Table 10 presents the differences between a_R^2 in the training phase and in the cross-validation and testing phases. The minimum sample size at which these differences fell below 0.05 was 104 observations. The exception was the RBF-NN algorithm, which reached this threshold with 78 observations, suggesting an advantage in attaining a stable fit with fewer data points.

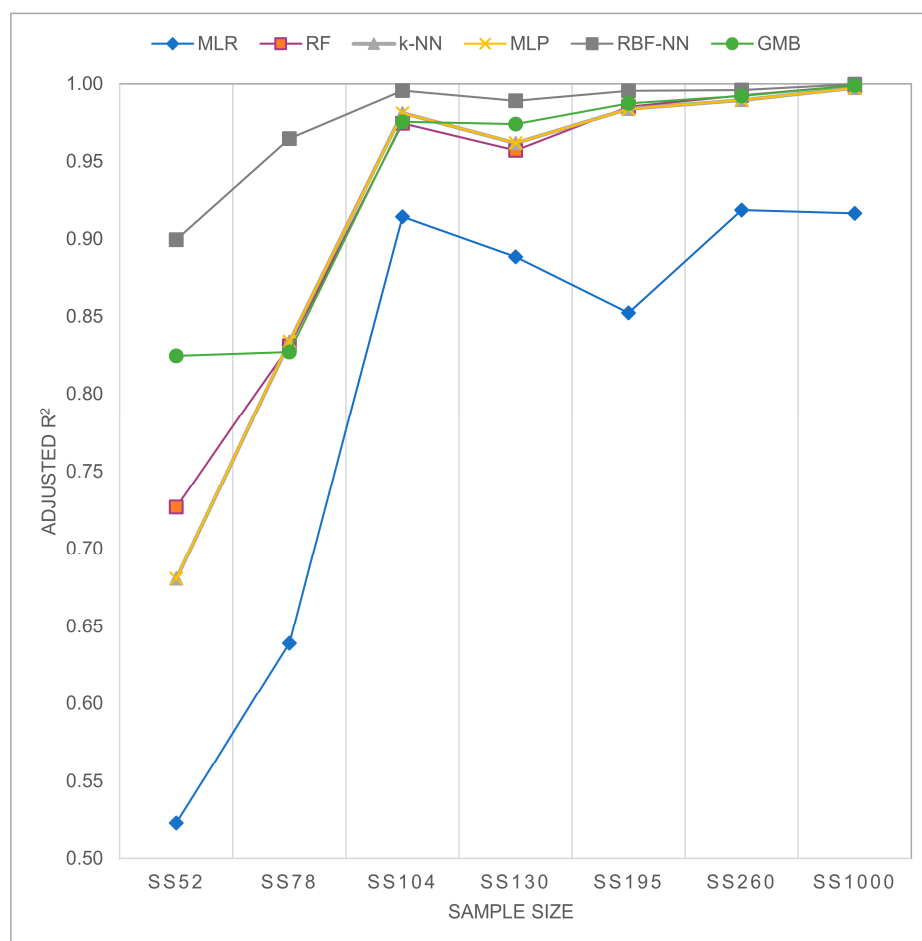


Figure 7. Comparison of a_{R^2} values in the test phase for models with different sample sizes.

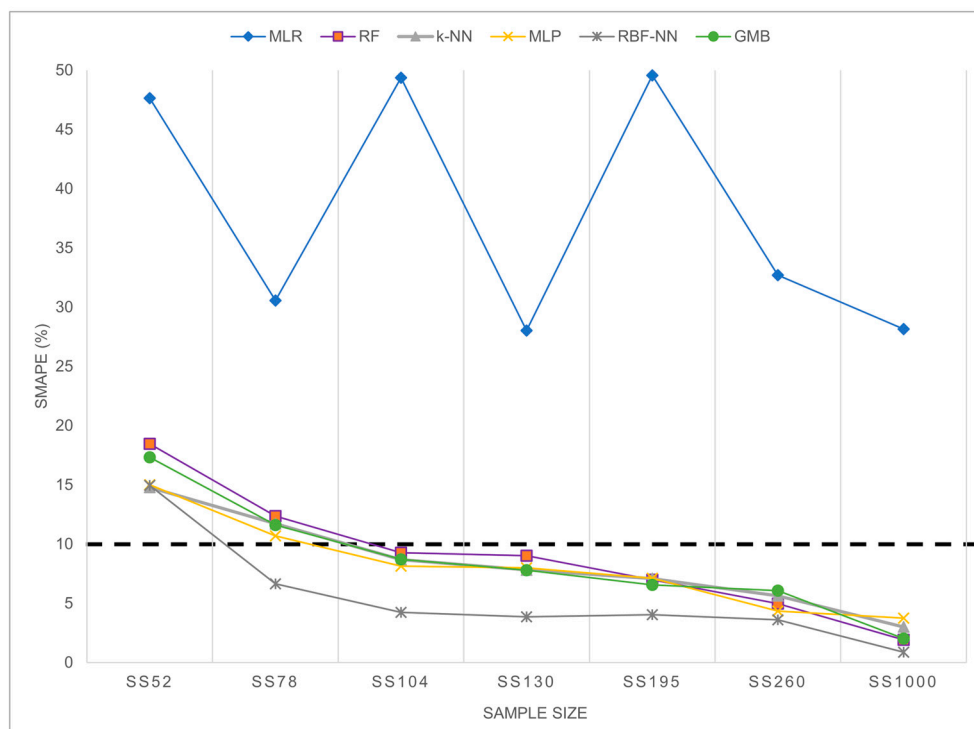


Figure 8. Comparison of SMAPEs in the testing phase for algorithms with different sample sizes.

Table 10. Difference between the a_R^2 values for the training phase and those for the cross-validation and test phases. The acronyms ‘Train’, ‘Val’, and ‘Test’ denote training, validation, and testing, respectively.

Model	Difference (train_val)	Difference (train_test)	Model	Difference (train_val)	Difference (train_test)
MLR52	0.24	0.39	MLP52	0.05	0.10
MLR78	0.23	0.29	MLP78	0.02	0.08
MLR104	0.04	0.02	MLP104	0.01	0.01
MLR130	0.05	0.02	MLP130	0.01	0.00
MLR195	0.06	0.06	MLP195	0.01	0.01
MLR260	0.02	0.01	MLP260	0.00	0.00
MLR1000	0.00	0.00	MLP1000	0.00	0.00
RF52	0.18	0.26	RBF-NN 52	0.01	0.09
RF78	0.12	0.17	RBF-NN78	0.01	0.03
RF104	0.03	0.02	RBF-NN104	0.00	0.00
RF130	0.03	0.04	RBF-NN130	0.01	0.01
RF195	0.01	0.01	RBF-NN195	0.00	0.00
RF260	0.01	0.01	RBF-NN 260	0.00	0.00
RF1000	0.00	0.00	RBF-NN 1000	0.00	0.00
k-NN52	0.10	0.32	GMB52	0.12	0.18
k-NN78	0.03	0.17	GMB78	0.05	0.17
k-NN104	0.03	0.02	GMB 104	0.04	0.02
k-NN130	0.03	0.04	GMB130	0.02	0.03
k-NN195	0.01	0.02	GMB195	0.01	0.01
k-NN260	0.01	0.01	GMB260	0.01	0.01
k-NN1000	0.01	0.00	GMB1000	0.00	0.00

3.7.2. Goodness of Fit and Accuracy of the Models (Condition 2)

Of the six models that met Condition 1, five also met Condition 2, achieving an $a_R^2 > 0.8$ and an SMAPE $< 10\%$ in the testing phase. Likewise, these five models exhibited a testing RMSE lower than the standard deviation of the training sample, indicating that prediction errors are smaller than the natural variability of the data. Consequently, the MLR104 model was excluded from the selection process, as its testing SMAPE exceeded 10%, as shown in Table 11.

Table 11. a_R^2 , RMSE, and SMAPE in the model testing phase and the standard deviation of the sample used for training.

Model	a_R^2	Test Phase RMSE	Test Phase SMAPE	SD
MLR104	0.9144	0.016	49.3	0.057
RF104	0.9745	0.0081	9.3	0.057
k-NN104	0.981	0.007	8.7	0.057
MLP104	0.989	0.043	8.1	0.057
RBF-NN78	0.965	0.0044	6.6	0.0495
GMB104	0.975	0.008	8.7	0.057

3.7.3. External Validation (Condition 3)

In Figure 9, the red hatched bars show, for the six models that met Condition 2, the SMAPE (%) computed between the prediction surface and the in-situ measurements, which was $< 10\%$ in all cases. The blue hatched bars represent the SMAPEs in the testing phase. The comparison indicates similar accuracy in both scenarios, except for k-NN 104 and GMB104, whose SMAPEs are reduced by approximately half with in situ data, indicating better field predictions and, consequently, more accurate prediction surfaces.

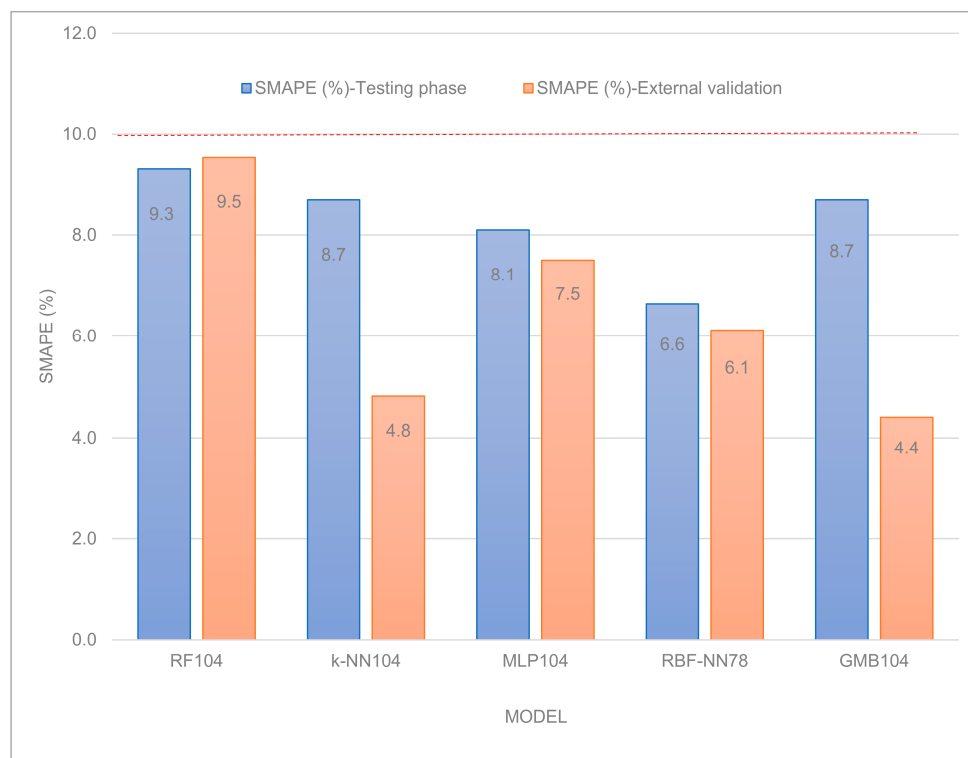


Figure 9. Comparison of SMAPEs (%) calculated in the testing phase and in external validation.

The evaluation of the three conditions determined that the minimum sample size with the best performance was 104 observations. Only the RBF-NN model met the conditions with 78 observations.

Figure 10 compares a_R^2 across the training, validation, and testing phases for the five models that satisfied the three conditions. All five models show very similar a_R^2 values (>0.90) across all phases, indicating high stability and explanatory power.

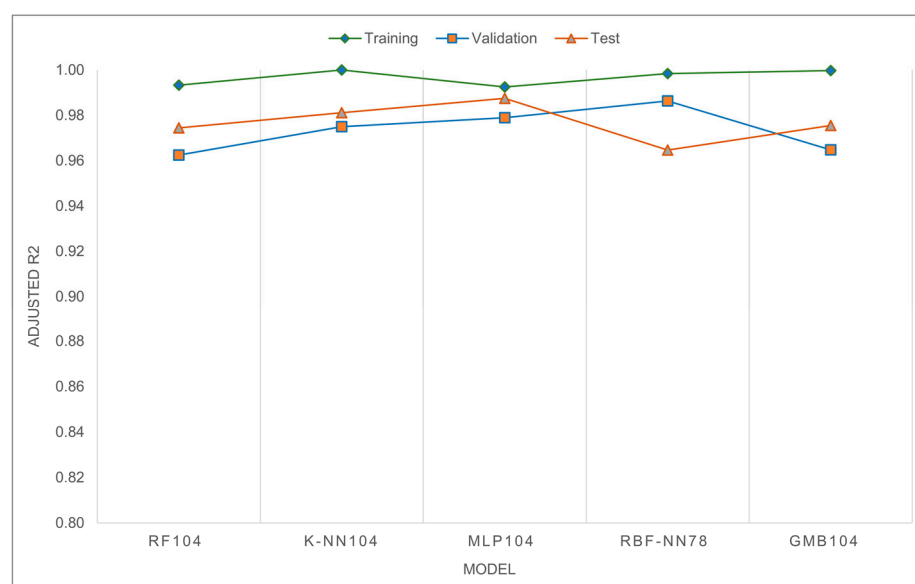


Figure 10. Comparison of a_R^2 across the training, validation, and testing phases for the five models that satisfied the three conditions.

Based on these findings, RBF-NN delivered the best performance with a small training sample, achieving suitable metrics with 78 observations. Among the models trained with

104 observations, GMB performed best, recording the lowest SMAPE (4.4%) based on in situ data.

3.8. Prediction Surfaces

Figure 11 displays the prediction surfaces and their basic statistics, derived from the five models that satisfied the three conditions. The strong visual and statistical similarity supports the consistency of using these algorithms to predict NH_3 surfaces.

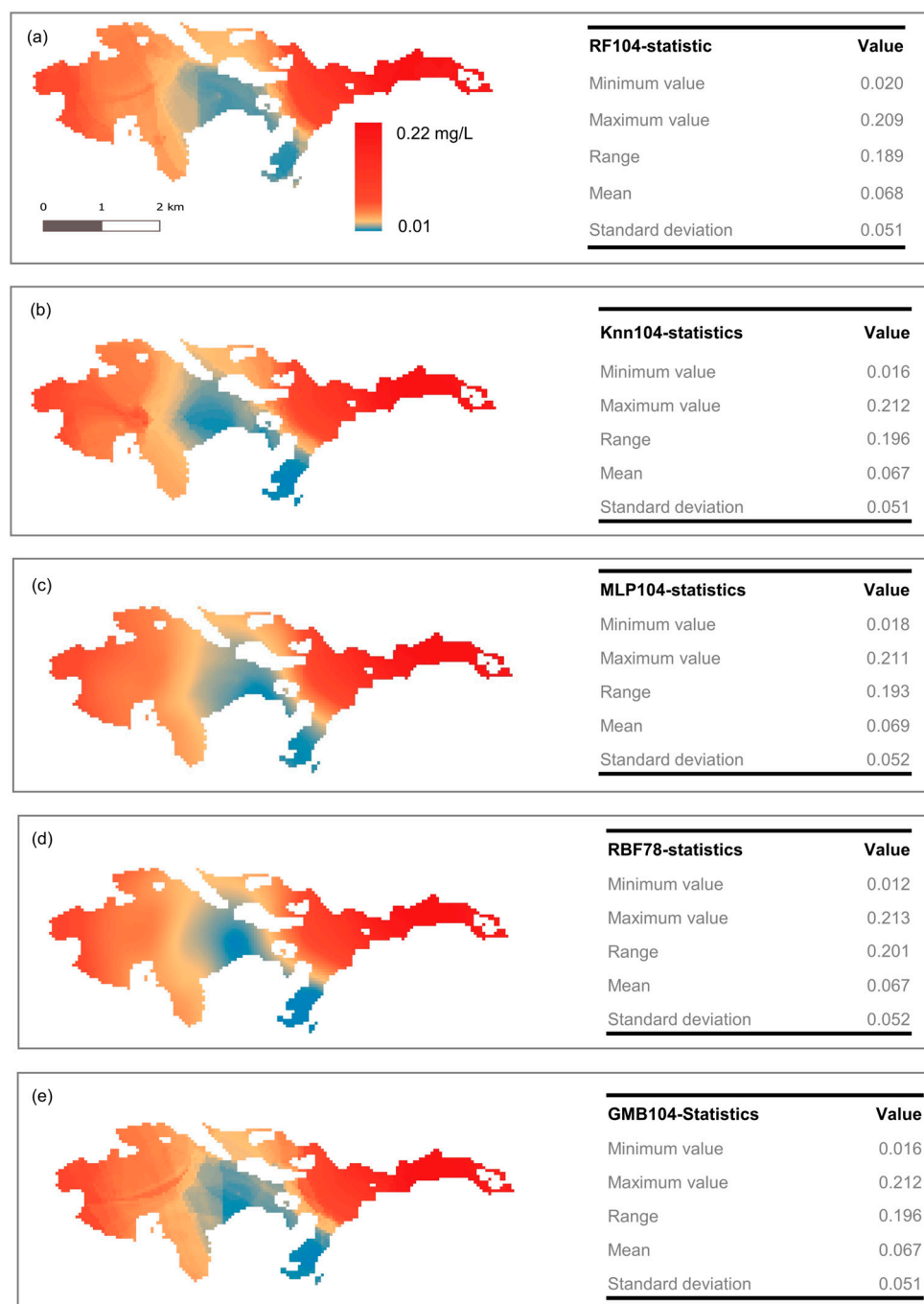


Figure 11. Prediction surfaces and their basic statistics derived from the five models that met the three conditions: (a) NH_3 -RF104; (b) NH_3 -k-NN104; (c) NH_3 -MLP104; (d) NH_3 -RBF-NN78; and (e) NH_3 -GMB104.

Table 12 presents the spatial correlation matrix between the six prediction surfaces. A high degree of spatial correlation (values > 0.90) is observed, indicating that the NH₃ concentration predictions obtained with any of the six models are similar.

Table 12. A spatial correlation matrix between the six prediction surfaces derived from the five models that met the three conditions.

RF104	k-NN104	MLP104	RBF-NN78	GMB104	
1.00	0.99	1.00	0.99	0.99	RF104
	1.00	1.00	1.00	1.00	k-NN104
		1.00	1.00	1.00	MLP104
			1.00	1.00	RBF-NN78
				1.00	GMB104

4. Discussion

This research is situated within the data-centric artificial intelligence approach, addressing the paucity of in situ observations for water quality modelling—a critical constraint for AI-based modelling of seasonal hydrological systems. The primary objective was to identify the minimum sample size required to train robust machine learning models capable of accurately predicting water quality. To increase data availability, ordinary and EBK were applied to a sample of in situ water quality measurements.

Based on an in situ sample of 30 observations that were collected on 27 September 2024 during the rainy season, data augmentation was performed using OK and EBK. A subset of 1000 points from the augmented sample, spatially well distributed across the Guartínaja wetland within a complex of tropical wetlands in northern Colombia, was selected. The NH₃ concentration was adopted as the target variable, while TDS, SS, PO₄, DO, NO₃ and COD served as predictor variables following a dimensionality reduction analysis.

Six ML algorithms (MLR, RF, k-NN, GMB, MLP and RBF-NN) were developed. They were initially trained on the 1000-point sample to generate learning curves, which indicated that model performance stabilised at around 200 data points. The augmented dataset was divided into subsamples of varying sizes (26, 52, 78, 104, 130, 195 and 260 points), from which the six ML models were trained to determine the minimum sample size required for robust water quality modelling. Five models (RF, k-NN, MLP, RBF-NN and GMB) achieved high explanatory power for NH₃, with $a_{R^2} > 0.95$ and SMAPEs < 10% in the test phase. Similar results were observed for the prediction surfaces.

The models' evaluation under the criteria of stability, goodness of fit and external validation of the models showed that the optimal minimum sample size for most algorithms was 104 data points.

4.1. Data Augmentation

The use of the two interpolation methods, OK and EBK, did not constitute a comparative analysis; the method that best fitted the in situ sample was selected. The kriging surfaces generated for the different parameters effectively reproduced the variability of the original sample, as the RMSE from cross-validation was lower than the SD of the original data.

We conclude that OK and EBK are appropriate for augmenting datasets used to train AI models for water quality applications. Validation of the kriging-derived prediction surfaces against in situ measurements yielded SMAPEs below 10% and an RMSE, supporting the methods' suitability. The application of geostatistical methods for data augmentation is a recent development documented by a limited number of authors. Ref. [17] employed two interpolation techniques to augment an agricultural database: Gaussian processes

(GP) with variable kernels and kriging with various variograms. They found that the GP method outperformed kriging in predictive performance, whereas kriging produced more homogeneous spatial coverage [18]. This likewise reported favourable results for kriging: data generated by kriging were used to train a CNN–RF model and were compared with a model trained without pseudo-labels, demonstrating that the inclusion of pseudo-labelled images substantially improved prediction error and spatial correlation on the test data.

Ref. [53] trained a random forest regression model to predict barium (Ba) concentration (in mg/kg) using different datasets: (a) kriging-derived data; (b) field-captured data; (c) unlabelled spatial data, treated as additional covariates; and (d) self-training data. On the other hand, ref. [54] also applied the kriging method to augment geomagnetism, altitude, and slope data. These researchers trained a random forest model to generate and spatialise geochemical content predictions. All authors, as in the present study, agreed that model statistics improved considerably when incorporating data derived from geostatistical methods.

In AI-based water quality modelling, the study by [16] is one of the published works that evaluated the effectiveness of OK for enhancing models that predict TDS. A statistically significant difference was found between models trained on field samples and those trained on augmented samples; consequently, OK was concluded to be an effective technique for enhancing model quality. Our findings are consistent with [16]: kriging produced surfaces that reproduced the variability of the field data, and the augmented data enabled models whose predictors account for over 90% of the variance in NH_3 .

Ref. [53] notes that geostatistical methods perform best when observations of the target variable exhibit appreciable spatial dependence; they are unsuitable when spatial autocorrelation is very weak or negligible. Furthermore, we emphasise that the effectiveness of geostatistical methods for data augmentation depends on a thorough evaluation of the in situ sample to determine its suitability for geostatistical treatment, as discussed [30,55]. Requirements for a sample to be suitable for interpolation include an approximately Gaussian distribution, demonstrable spatial autocorrelation, stationarity, a sufficient sample size and an appropriate spatial configuration that captures the phenomenon's variability. After constructing kriging surfaces but before using them for data augmentation, it is essential to perform cross-validation and statistical comparison between the original sample and the interpolated surface [56]. We attribute the excellent statistical performance of the prediction surfaces in this study to adherence to these protocols.

A considerable number of researchers have highlighted the lack of in situ data as a limitation in modelling [57–59]. It can be argued that the absence of sufficient data necessary for generating robust models becomes a barrier to the integration of artificial intelligence into environmental modelling. While initiatives employing geostatistical methods for data augmentation are relatively recent and limited in number, these methods could be seen as a potential solution to the deficiency of in situ spatial data. In an environmental context, this presents significant opportunities for researchers in regions lacking robust monitoring policies for environmental resources or advanced technologies for real-time monitoring. By overcoming data limitations, it will become more feasible to engage with artificial intelligence in the analysis of geospatial data.

In water quality modelling, researchers have progressed towards bio-optical modelling [33] by utilising remote sensors as a strategy to increase data availability and achieve improved spatial distribution. However, in situ data remain crucial in these models, as they are necessary for calibration [60]. This suggests that, at least in the context of water quality, in situ data—which require considerable human and economic effort to collect [4]—are still irreplaceable. Therefore, the ability to reliably augment in situ data artificially represents a transformative advancement in the monitoring and understanding of water, which is an

essential resource for life on Earth. An increased quantity of data enhances the potential to employ a wide range of artificial intelligence algorithms, thereby improving the accuracy of predictions.

Our research, alongside other studies related to the use of interpolators as data augmentation methods, represents the beginning of a comprehensive research trajectory that will contribute to the body of knowledge in geoscience artificial intelligence (GeoIA). Reference [30] describes approximately 21 geostatistical and 12 deterministic interpolators that should be assessed as methods for data augmentation. Furthermore, in the realm of water quality, it is essential to integrate these techniques with bio-optical models and evaluate them against additional physicochemical parameters.

4.2. Comparison of Sample Sizes for NH₃ Parameter Modelling Using Artificial Intelligence Models

The six machine learning algorithms (MLR, RF, k-NN, MLP, RBF-NN, and GBM) were initially trained using a sample comprising 1000 points derived from the data augmentation process. This training resulted in the generation of a learning curve, the analysis of which indicated that the maximum stabilisation of the six algorithms occurs with a sample size of 200 data points. For this sample size, all models achieved a training and validation performance greater than 0.9 (a_R^2). Similar findings were reported by [61], which assessed a random forest model to predict forest above-ground biomass at local, regional, and national scales, comparing various sample sizes ($n = 25, 50, 100, 200, 300$, and 500). They noted that a_R^2 values increased with larger sample sizes; however, beyond a minimum of 300 to 500 samples, the improvement in accuracy became negligible, highlighting the diminishing returns associated with the use of excessively large sample sizes.

Seven subsamples were extracted from the augmented sample, comprising sizes of 26, 52, 78, 104, 130, 195, and 260 data points. Each of the ML algorithms was trained using the eight datasets (including the augmented sample of 1000 data points), resulting in a total of 48 trained models. A consistent pattern of behaviour was observed across all algorithms, indicating that the models' generalisation ability improved as the sample size increased, a phenomenon also reported in various studies [21,61,62].

None of the algorithms trained with a subsample of 26 data points achieved an a_R^2 greater than 0.5 during the test phase. For the algorithms trained with a subsample of 52 data points, overfitting was evident in the MLR, RF, and k-NN models, findings that align with those of [61], which noted that no model exhibited consistent accuracy with small sample sizes ($n < 30$).

The models trained with a sample size of 104 data points demonstrated stability across their three phases: training, cross-validation, and testing, with the a_R^2 exceeding 0.95 in all phases. Regarding SMAPEs, the highest error was observed for the MLR algorithm, where the expected pattern of decreasing errors with increasing data was not realised, as all MLR models exhibited errors greater than 30%. Conversely, the algorithm showing the lowest error across all sample sizes was RBF-NN.

The analysis of the optimal sample size, conducted through the evaluation of the models under conditions of stability, adjustment capacity, and external validation, determined that the best minimum sample size for most algorithms was 104 data points. There were two exceptions: the MLR algorithm, which exhibited SMAPE values exceeding 10% in both the test phase and the external evaluation, and the RBF-NN algorithm, which fulfilled all three conditions using a sample size of 78 data points, indicating a superior explanatory capacity of the model with a smaller sample.

The consistency of these results is evident in the prediction surfaces derived from the selected models: RF104, k-NN 104, MLP104, RBF-NN78, and GMB104. The spatial correlation matrix of the prediction surfaces revealed very high correlation values (ranging from

0.99 to 1), indicating that any of these surfaces would be suitable for making predictions of the NH_3 parameter.

The findings described suggest that the ideal minimum sample size may vary depending on the algorithm and other factors that may not have been considered in this study. For instance, ref. [63] investigated how variations in training set size, ranging from a large sample ($n = 10,000$) to a very small one ($n = 40$), affect the performance of six supervised machine learning algorithms applied to the classification of high spatial resolution (AR) remote sensing data (1–5 m) over large areas. Overall, global accuracy improved with increasing sample size; for all classification methods, the highest overall mean accuracy was achieved with the 10,000-sample set, while the lowest was observed with the 40-sample set. However, each classifier responded differently to the increase in sample size.

Ref. [64] evaluated the effects of training sample size, sampling design, and prediction model type (MLR and RF) on the prediction performance for three soil properties: soil organic carbon, clay, and pH value. They also employed geostatistical techniques to obtain sufficient data. Their overall conclusion was that a larger sample size proved to be the most critical factor for accuracy within the low to medium sample size range. However, there were other findings worth highlighting: (a) although increasing the training sample size improved prediction accuracy, the relative improvement decreased progressively, and (b) the stabilisation of the improvement in accuracy was more pronounced with MLR than with RF. Therefore, random forest performed considerably better with larger sample sizes.

In machine learning, the algorithm is not the sole key factor influencing the accuracy of the final classification [10]; however, it cannot be asserted that this accuracy depends exclusively on the amount of data. The findings of this research, alongside other studies related to sample size for artificial intelligence models, suggest the necessity of considering additional factors in sample size analyses, such as the type of algorithm, various data augmentation methods, the target variable, and the spatial extent of the study area.

With regard to the target variable, it should be noted that water quality is a highly dynamic attribute, whose behaviour varies spatially and temporally, as well as between different parameters. In this study, we focus on ammonium (NH_3); therefore, the estimated sample size may differ for other indicators or for sampling that is carried out in different climatic seasons. Nevertheless, the methodology applied offers a replicable framework that can serve as a reference for studies on other water quality parameters.

5. Conclusions

This study demonstrates that interpolation via ordinary and EBK for data augmentation produces high-quality models for predicting the NH_3 water quality parameter. These findings represent a significant advance in addressing data scarcity in water quality modelling and are expected to facilitate broader application of artificial intelligence techniques in this domain.

Within the data-centric artificial intelligence framework, this research is particularly pertinent: it proposes a strategy that integrates statistical methods, geospatial technologies and artificial intelligence to produce more robust models with improved predictive accuracy. Enhanced predictions will support more precise diagnostics and decision-making for the management of natural aquatic ecosystems.

One limitation of this study is our exclusive focus on ammonium as a water quality parameter, which was based on a sample collected during the wet season. While this approach enabled us to conduct a comprehensive analysis and train 48 artificial intelligence models, it is probable that the minimum sample sizes would differ for other parameters and across various climatic seasons.

The methodology is applicable to other water quality parameters and to other quantitative, spatially continuous environmental variables, representing a notable contribution to the emerging field of GeoAI.

The findings, together with comparisons to related studies, indicate that research on AI applied to geoinformation—particularly in the context of water quality—remains at an early stage. Future research avenues include (1) evaluating additional interpolators as data augmentation methods and (2) applying this methodological framework to additional parameters and evaluating it in different seasons to assess its generality because spatial continuity varies among water quality parameters and across climatic seasons.

Author Contributions: D.M.A.: conceptualisation, formal analysis, investigation, resources, software, validation, visualisation, writing—original draft, writing—review and editing. F.T.B.: field measurements, data curation, writing—review and editing. V.S.B.: writing—review and editing, graphic design, language review. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Science and Technology of Colombia. Grant number: 2022-0786.

Data Availability Statement: The data presented in the study are available upon request from the first and corresponding author. The data are not publicly available due to privacy.

Acknowledgments: We thank the University of Córdoba for facilitating researchers' participation in science and technology projects and for providing logistical support for their implementation.

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

Abbreviations

The following abbreviations are used in this manuscript:

a_R2	Adjusted Coefficient of Determination
AI	Artificial Intelligence
BAO	Biological Availability of Oxygen
BSWC	Bajo Sinú Wetland Complex
CNN-RF	Convolutional Neural Network–Random Forest
CopulaGAN	Copula Generative Adversarial Network
CTGAN	Conditional Tabular Generative Adversarial Network
CV	Coefficient of Variation
DO	Dissolved Oxygen
DOC	Chemical Availability of Oxygen
EC	Electrical Conductivity
EBK	Empirical Bayesian Kriging
GBM	Gradient Boosting Machines
GC	Gaussian Copula
GeoAI	Geospatial Artificial Intelligence
GMB1000	GMB Algorithm Trained Using a Sample of 1000 Data Points
GMB104	GMB Algorithm Trained Using a Sample of 104 Data Points
GMB130	GMB Algorithm Trained Using a Sample of 130 Data Points
GMB195	GMB Algorithm Trained Using a Sample of 195 Data Points
GMB26	GMB Algorithm Trained Using a Sample of 26 Data Points
GMB260	GMB Algorithm Trained Using a Sample of 260 Data Points
GMB52	GMB Algorithm Trained Using a Sample of 52 Data Points
GMB78	GMB Algorithm Trained Using a Sample of 78 Data Points
GP	Gaussian Processes
I _{moran}	Global Moran's I index

IoT	Internet of Things
k-NN	k-Nearest Neighbours
K-NN1000	K-NN Algorithm Trained Using a Sample of 1000 Data Points
K-NN104	K-NN Algorithm Trained Using a Sample of 104 Data Points
K-NN130	K-NN Algorithm Trained Using a Sample of 130 Data Points
K-NN195	K-NN Algorithm Trained Using a Sample of 195 Data Points
K-NN26	K-NN Algorithm Trained Using a Sample of 26 Data Points
K-NN260	K-NN Algorithm Trained Using a Sample of 260 Data Points
K-NN52	K-NN Algorithm Trained Using a Sample of 52 Data Points
K-NN78	K-NN Algorithm Trained Using a Sample of 78 Data Points
MLP	Multilayer Perceptron regressor
MLP1000	MLP Algorithm Trained Using a Sample of 1000 Data Points
MLP104	MLP Algorithm Trained Using a Sample of 104 Data Points
MLP130	MLP Algorithm Trained Using a Sample of 130 Data Points
MLP195	MLP Algorithm Trained Using a Sample of 195 Data Points
MLP26	MLP Algorithm Trained Using a Sample of 26 Data Points
MLP260	MLP Algorithm Trained Using a Sample of 260 Data Points
MLP52	MLP Algorithm Trained Using a Sample of 52 Data Points
MLP78	MLP Algorithm Trained Using a Sample of 78 Data Points
MLR	Multiple Linear Regression
MLR1000	MLR Algorithm Trained Using a Sample of 1000 Data Points
MLR104	MLR Algorithm Trained Using a Sample of 104 Data Points
MLR130	MLR Algorithm Trained Using a Sample of 130 Data Points
MLR195	MLR Algorithm Trained Using a Sample of 195 Data Points
MLR26	MLR Algorithm Trained Using a Sample of 26 Data Points
MLR260	MLR Algorithm Trained Using a Sample of 260 Data Points
MLR52	MLR Algorithm Trained Using a Sample of 52 Data Points
MLR78	MLR Algorithm Trained Using a Sample of 78 Data Points
NH ₃	Ammonia
NH ₃	Ammonium
NH ₃ _104	Sample of NH ₃ Containing 104 Data Points
NH ₃ _130	Sample of NH ₃ Containing 130 Data Points
NH ₃ _195	Sample of NH ₃ Containing 195 Data Points
NH ₃ _26	Sample of NH ₃ Containing 26 Data Points
NH ₃ _260	Sample of NH ₃ Containing 260 Data Points
NH ₃ _52	Sample of NH ₃ Containing 52 Data Points
NH ₃ _78	Sample of NH ₃ Containing 78 Data Points
NO ₂	Nitrite Ion
NO ₃	Nitrate Ion
NORG	Organic Nitrogen
OK	Ordinary Kriging
OLS	Ordinary Least Squares
p	Pearson Correlation
P T	Total Phosphorus
pH	Hydrogen Ion Potential
PO	Phosphate
p-value	Probability
R2	Coefficient of Determination
RBF-NN	Radial Basis Function Neural Network.
RBF-NN1000	RBF-NN Algorithm Trained Using a Sample of 1000 Data Points
RBF-NN104	RBF-NN Algorithm Trained Using a Sample of 104 Data Points
RBF-NN130	RBF-NN Algorithm Trained Using a Sample of 130 Data Points
RBF-NN195	RBF-NN Algorithm Trained Using a Sample of 195 Data Points
RBF-NN26	RBF-NN Algorithm Trained Using a Sample of 26 Data Points

RBF-NN260	RBF-NN Algorithm Trained Using a Sample of 260 Data Points
RBF-NN52	RBF-NN Algorithm Trained Using a Sample of 52 Data Points
RBF-NN78	RBF-NN Algorithm Trained Using a Sample of 78 Data Points
RF	Random Forest
RF1000	RF Algorithm Trained Using a Sample of 1000 Data Points
RF104	RF Algorithm Trained Using a Sample of 104 Data Points
RF130	RF Algorithm Trained Using a Sample of 130 Data Points
RF195	RF Algorithm Trained Using a Sample of 195 Data Points
RF26	RF Algorithm Trained Using a Sample of 26 Data Points
RF260	RF Algorithm Trained Using a Sample of 260 Data Points
RF52	RF Algorithm Trained Using a Sample of 52 Data Points
RF78	RF Algorithm Trained Using a Sample of 78 Data Points
RMSE	Root Mean Squared Error
SAL	Salts
SD	Standard Deviation
SMAPE	Symmetric Mean Absolute Percentage Error
SS	Suspended Solids
SS	Sample Size
SS1000	Sample Size of 1000 Data Points
SS104	Sample Size of 104 Data Points
SS130	Sample Size of 130 Data Points
SS195	Sample Size of 195 Data Points
SS260	Sample Size of 260 Data Points
SS52	Sample Size of 52 Data Points
SS78	Sample Size of 78 Data Points
SST	Total Suspended Solids
T	Temperature
TDS	Total Dissolved Solids
Test	Test
Train	Training
TURB	Turbidity
TVAE	Tabular Variational Autoencoder
Val	Cross-validation
VIF	Variance Inflation Factor

Appendix A

Coordinate reference system (CRS)

Name

EPSG:4326—WGS 84

Units

Geographic (uses latitude and longitude for coordinates)

Type

Geographic (2D)

Method

Lat/long (Geodetic alias)

Celestial Body

Earth

Accuracy

Based on the World Geodetic System 1984 ensemble (EPSG:6326), which has a limited pre-accuracy of 2 m at best.

ID	POINT_X	POINT_Y
GUA 4	−75.725662	9.223307
GUA 5	−75.737405	9.222205
GUA 6	−75.741173	9.214303
GUA 7	−75.751473	9.220363
GUA 8	−75.758172	9.215152
GUA 9	−75.761435	9.224305
GUA 10	−75.774813	9.220948
GUA 12	−75.762158	9.227752
GUA 13	−75.76525	9.21904
GUA 15	−75.745938	9.22572
GUA 16	−75.742662	9.219945
GUA 17	−75.77354806	9.222718444
GUA 18	−75.76850539	9.220549554
GUA 19	−75.76812583	9.216645551
GUA 20	−75.75874538	9.221037554
GUA 21	−75.76508938	9.228194892
GUA 22	−75.75500404	9.219302442
GUA 23	−75.75711871	9.210843769
GUA 24	−75.73879159	9.218597552
GUA 25	−75.73147158	9.222176222
GUA 26	−75.72805558	9.224399334
GUA 27	−75.7524556	9.227327336
GUA 28	−75.75484138	9.227978003
GUA 29	−75.74315987	9.211006436
GUA 30	−75.76578072	9.223714778
GUA 1	−75.74129598	9.224562001
GUA 2	−75.75680016	9.224477279
GUA 3	−75.72240291	9.223545334
GUA 11	−75.76781311	9.22550905
GUA 14	−75.75359282	9.225971859

Appendix B

Table A1. Compares the RMSE from cross-validation of each kriging method with the standard deviation of the in situ sample.

Parameter	RMSE (Cross Validation)	SD (In Situ Sample)
NH ₃	0.04	0.06
COD	4.8	8.01
DO	0.5	0.55
NO ₃	0.11	0.29
SS	0.23	0.22
TDS	2.1	6.89
PO ₄	0.04	0.03

Table A2. Provides the results of OK interpolation for the parameters COD, DO, NO₃, SS, and NH₃.

Item	COD	DO	NO ₃	SS	NH ₃
Records	30	30	30	30	30
Neighbours	smooth	5	Smooth	Smooth	Smooth
Number of lags	12	12	12	12	12
Model type	Stable	Stable	Stable	Stable	Stable
Range	0.0163	0.018	0.016	0.022	0.017
Anisotropy	Yes	Yes	Yes	Yes	Yes
Direction	114.6	114.6	138.3	0	28.8
Partial sill	84.49	84.49	0.07	0	0.48
Transformation	None	None	BoxCo	None	Log

Table A3. Provides the results of EBK interpolation for the parameters PO₄ and TDS.

Item	TDS	PO ₄
Records	30	30
Transformation	None	None
Semivariogram Model	Power	Power
Type		
Subset Size	100	100
Overlap Factor	1	1
Number Simulations	100	100
Neighbours to include	15	Smooth
Include at least	10	

Appendix C

Tables A4–A8 present the hyperparameters utilised for training each of the 48 machine learning models.

Table A4. Hyperparameters utilised for training the RF algorithm models.

Data Division	Best Hyperparameters
RF26 (20, 6)	{‘max_depth’: 5, ‘max_features’: ‘sqrt’, ‘min_samples_leaf’: 1, ‘min_samples_split’: 2, ‘n_estimators’: 200}
RF 52(40, 12)	{‘max_depth’: None, ‘max_features’: ‘sqrt’, ‘min_samples_leaf’: 1, ‘min_samples_split’: 2, ‘n_estimators’: 100}
RF 78 (60, 18)	{‘max_depth’: None, ‘max_features’: ‘sqrt’, ‘min_samples_leaf’: 1, ‘min_samples_split’: 2, ‘n_estimators’: 50}
RF 104 (80, 24)	{‘max_depth’: None, ‘max_features’: ‘sqrt’, ‘min_samples_leaf’: 1, ‘min_samples_split’: 2, ‘n_estimators’: 200}
RF 130 (100, 30)	{‘max_depth’: None, ‘max_features’: ‘sqrt’, ‘min_samples_leaf’: 1, ‘min_samples_split’: 2, ‘n_estimators’: 200}
RF 195 (150, 45)	{‘max_depth’: None, ‘max_features’: ‘sqrt’, ‘min_samples_leaf’: 1, ‘min_samples_split’: 2, ‘n_estimators’: 200}

Table A4. *Cont.*

Data Division	Best Hyperparameters
RF 260 (200, 60)	{‘max_depth’: None, ‘max_features’: ‘sqrt’, ‘min_samples_leaf’: 1, ‘min_samples_split’: 2, ‘n_estimators’: 200}
RF 1000 (600, 400)	{‘max_depth’: 20, ‘max_features’: ‘sqrt’, ‘min_samples_leaf’: 1, ‘min_samples_split’: 2, ‘n_estimators’: 50}

Table A5. Hyperparameters utilised for training the K-NN algorithm models.

Data Division	Best Hyperparameters
K-NN26 (20, 6)	{‘algorithm’: ‘auto’, ‘leaf_size’: 30, ‘metric’: ‘manhattan’, ‘n_neighbors’: 3, ‘p’: 1, ‘weights’: ‘distance’}
K-NN 52(40, 12)	{‘algorithm’: ‘auto’, ‘leaf_size’: 30, ‘metric’: ‘manhattan’, ‘n_neighbors’: 3, ‘p’: 1, ‘weights’: ‘distance’}
K-NN 78 (60, 18)	{‘algorithm’: ‘auto’, ‘leaf_size’: 30, ‘metric’: ‘manhattan’, ‘n_neighbors’: 3, ‘p’: 1, ‘weights’: ‘distance’}
K-NN 104 (80, 24)	{‘algorithm’: ‘auto’, ‘leaf_size’: 30, ‘metric’: ‘euclidean’, ‘n_neighbors’: 3, ‘p’: 1, ‘weights’: ‘distance’}
K-NN 130 (100, 30)	{‘algorithm’: ‘auto’, ‘leaf_size’: 30, ‘metric’: ‘manhattan’, ‘n_neighbors’: 3, ‘p’: 1, ‘weights’: ‘distance’}
K-NN 195 (150, 45)	{‘algorithm’: ‘auto’, ‘leaf_size’: 30, ‘metric’: ‘manhattan’, ‘n_neighbors’: 3, ‘p’: 1, ‘weights’: ‘distance’}
K-NN 260 (200, 60)	{‘algorithm’: ‘auto’, ‘leaf_size’: 30, ‘metric’: ‘euclidean’, ‘n_neighbors’: 3, ‘p’: 1, ‘weights’: ‘distance’}
K-NN 1000 (600, 400)	{‘algorithm’: ‘auto’, ‘leaf_size’: 30, ‘metric’: ‘euclidean’, ‘n_neighbors’: 3, ‘p’: 1, ‘weights’: ‘distance’}

Table A6. Hyperparameters utilised for training the MLP algorithm models.

Data Division	Best Hyperparameters
MLP26 (20, 6)	{‘activation’: ‘relu’, ‘alpha’: 0.001, ‘hidden_layer_sizes’: (200, 100), ‘learning_rate’: ‘constant’, ‘max_iter’: 50, ‘solver’: ‘adam’}
MLP 52 (40, 12)	{‘activation’: ‘relu’, ‘alpha’: 0.001, ‘hidden_layer_sizes’: (100, 200), ‘learning_rate’: ‘constant’, ‘max_iter’: 50, ‘solver’: ‘adam’}

Table A6. *Cont.*

Data Division	Best Hyperparameters
MLP 78 (60,18)	{'activation': 'relu', 'alpha': 0.01, 'hidden_layer_sizes': (100, 200), 'learning_rate': 'constant', 'max_iter': 100, 'solver': 'adam'}
MLP 104 (80, 24)	{'activation': 'relu', 'alpha': 0.01, 'hidden_layer_sizes': (200, 100), 'learning_rate': 'constant', 'max_iter': 200, 'solver': 'adam'}
MLP 130 (100, 30)	{'activation': 'relu', 'alpha': 0.01, 'hidden_layer_sizes': (200, 100), 'learning_rate': 'constant', 'max_iter': 200, 'solver': 'adam'}
MLP 195 (150, 45)	{'activation': 'relu', 'alpha': 0.01, 'hidden_layer_sizes': (200, 100), 'learning_rate': 'constant', 'max_iter': 200, 'solver': 'adam'}
MLP 260 (200, 60)	{'activation': 'relu', 'alpha': 0.01, 'hidden_layer_sizes': (200, 100), 'learning_rate': 'constant', 'max_iter': 200, 'solver': 'adam'}
MLP 1000 (600, 400)	{'activation': 'relu', 'alpha': 0.01, 'hidden_layer_sizes': (100, 100), 'learning_rate': 'constant', 'max_iter': 200, 'solver': 'adam'}

Table A7. Hyperparameters utilised for training the RBF-NN algorithm models.

Data Division	Best Hyperparameters
RBF26 (20, 6)	{'num_centers': 10, 'sigma': 1.0}
RBF 52(40, 12)	{'num_centers': 5, 'sigma': 1.0}
RBF 78 (60, 18)	{'num_centers': 15, 'sigma': 1.0}
RBF 104 (80, 24)	{'num_centers': 15, 'sigma': 1.0}
RBF 130 (100, 30)	{'num_centers': 15, 'sigma': 1.0}
RBF 195 (150, 45)	{'num_centers': 15, 'sigma': 1.0}
RBF 260 (200, 60)	{'num_centers': 15, 'sigma': 1.0}
RBF 1000 (600, 400)	{'num_centers': 15, 'sigma': 1.0}

Table A8. Hyperparameters utilised for training the GMB algorithm models.

Data Division	Best Hyperparameters
GBM 52(40, 12)	{'learning_rate': 0.1, 'max_depth': 5, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 300, 'subsample': 0.7}
GBM 78 (60, 18)	{'learning_rate': 0.1, 'max_depth': 5, 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 300, 'subsample': 0.7}

Table A8. Cont.

Data Division	Best Hyperparameters
GBM 104 (80, 24)	{‘learning_rate’: 0.2, ‘max_depth’: 5, ‘min_samples_leaf’: 1, ‘min_samples_split’: 5, ‘n_estimators’: 300, ‘subsample’: 0.7}
GBM 130 (100, 30)	{‘learning_rate’: 0.1, ‘max_depth’: 5, ‘min_samples_leaf’: 4, ‘min_samples_split’: 2, ‘n_estimators’: 300, ‘subsample’: 0.8}
GBM 195 (150, 45)	{‘learning_rate’: 0.1, ‘max_depth’: 3, ‘min_samples_leaf’: 1, ‘min_samples_split’: 2, ‘n_estimators’: 300, ‘subsample’: 1.0}
GBM 260 (200, 60)	{‘learning_rate’: 0.1, ‘max_depth’: 5, ‘min_samples_leaf’: 2, ‘min_samples_split’: 10, ‘n_estimators’: 300, ‘subsample’: 0.7}
GBM 1000 (600, 400)	{‘learning_rate’: 0.1, ‘max_depth’: 5, ‘min_samples_leaf’: 1, ‘min_samples_split’: 5, ‘n_estimators’: 300, ‘subsample’: 0.7}

References

1. Biswas, A.; Sarkar, S.; Das, S.; Dutta, S.; Choudhury, M.R.; Giri, A.; Bera, B.; Bag, K.; Mukherjee, B.; Banerjee, K.; et al. Water scarcity: A global hindrance to sustainable development and agricultural production—A critical review of the impacts and adaptation strategies. *Camb. Prism. Water* **2025**, *3*, e4. [\[CrossRef\]](#)
2. Frincu, R.M. Artificial intelligence in water quality monitoring: A review of water quality assessment applications. *Water Qual. Res. J.* **2025**, *60*, 164–176. [\[CrossRef\]](#)
3. Jackson-Blake, L.A.; Sample, J.E.; Wade, A.J.; Helliwell, R.C.; Skeffington, R.A. Are our dynamic water quality models too complex? A comparison of a new parsimonious phosphorus model, S imply P, and INCA-P. *Water Resour. Res.* **2017**, *53*, 5382–5399. [\[CrossRef\]](#)
4. Alotaibi, E.; Nassif, N. Artificial intelligence in environmental monitoring: In-depth analysis. *Discov. Artif. Intell.* **2024**, *4*, 84. [\[CrossRef\]](#)
5. Bianchi, O.; Putro, H.P. Artificial intelligence in environmental monitoring: Predicting and managing climate change impacts. *Int. Trans. Artif. Intell.* **2024**, *3*, 85–96. [\[CrossRef\]](#)
6. Uddin, M.G.; Nash, S.; Olbert, A.I. A review of water quality index models and their use for assessing surface water quality. *Ecol. Indic.* **2021**, *122*, 107218. [\[CrossRef\]](#)
7. Chia, M.Y.; Koo, C.H.; Huang, Y.F.; Di Chan, W.; Pang, J.Y. Artificial intelligence generated synthetic datasets as the remedy for data scarcity in water quality index estimation. *Water Resour. Manag.* **2023**, *37*, 6183–6198. [\[CrossRef\]](#)
8. Kokol, P.; Kokol, M.; Zagoranski, S. Machine learning on small size samples: A synthetic knowledge synthesis. *Sci. Prog.* **2022**, *105*, 368504211029777. [\[CrossRef\]](#)
9. Jakubik, J.; Vössing, M.; Kühn, N.; Walk, J.; Satzger, G. Data-centric artificial intelligence. *Bus. Inf. Syst. Eng.* **2024**, *66*, 507–515. [\[CrossRef\]](#)
10. Pothuganti, S. Review on over-fitting and under-fitting problems in Machine Learning and solutions. *Int. J. Adv. Res. Electr. Electron. Instrum. Eng.* **2018**, *7*, 3692–3695. [\[CrossRef\]](#)
11. Zha, D.; Bhat, Z.P.; Lai, K.H.; Yang, F.; Jiang, Z.; Zhong, S.; Hu, X. Data-centric artificial intelligence: A survey. *ACM Comput. Surv.* **2025**, *57*, 1–42. [\[CrossRef\]](#)
12. Strickland, E. Andrew Ng: Not Great AI. 2022. Available online: <https://spectrum.ieee.org/andrew-ng-data-centric-ai> (accessed on 20 October 2025).
13. Khosla, C.; Saini, B.S. Enhancing performance of deep learning models with different data augmentation techniques: A survey. In *2020 International Conference on Intelligent Engineering and Management (ICIEM)*; IEEE: New York, NY, USA, 2020; pp. 79–85. [\[CrossRef\]](#)

14. Maharana, K.; Mondal, S.; Nemade, B. A review: Data pre-processing and data augmentation techniques. *Glob. Transit. Proc.* **2022**, *3*, 91–99. [\[CrossRef\]](#)
15. Wang, Z.; Wang, P.; Liu, K.; Wang, P.; Fu, Y.; Lu, C.T.; Aggarwal, C.C.; Pei, J.; Zhou, Y. A Comprehensive Survey on Data Augmentation. *IEEE Trans. Knowl. Data Eng.* **2026**, *38*, 47–66. [\[CrossRef\]](#)
16. Ávila, D.M.; Bejarano, F.T.; Barrera, V.C.S. Ordinary kriging as a data augmentation technique for enhanced total dissolved solids prediction in tropical wetlands using Sentinel-2 and machine learning. *Int. J. Remote Sens.* **2026**, *47*, 1727–1758. [\[CrossRef\]](#)
17. Ferber, F.F.; Gay, D.; Soulié, J.C.; Diatta, J.; Maillard, O.A. Kriging and Gaussian Process Interpolation for Georeferenced Data Augmentation. *arXiv* **2025**, arXiv:2501.07183. [\[CrossRef\]](#)
18. Duan, L.; Jiang, Z.; Carlson, D. Augmenting Ground-Level PM_{2.5} Prediction via Kriging-Based Pseudo-Label Generation. *arXiv* **2024**, arXiv:2401.08061.
19. Kodenko, M.R.; Bobrovskaya, T.Y.M.; Reshetnikov, R.V.; Arzamasov, K.M.; Vladzymysky, A.V.; Omelyanskaya, O.G.V.E.; Vasilev, Y.A. Empirical approach to sample size estimation for testing of AI algorithms. In *Doklady Mathematics*; Pleiades Publishing: Moscow, Russia, 2024; Volume 110, pp. S62–S74. [\[CrossRef\]](#)
20. Kim, J.H.; Kwon, N.; Pandis, N.; Lee, S.J. Sample size calculation for an artificial intelligence study. *Am. J. Orthod. Dentofac. Orthop.* **2025**, *167*, 616–620. [\[CrossRef\]](#) [\[PubMed\]](#)
21. Moghaddam, D.D.; Rahmati, O.; Panahi, M.; Tiefenbacher, J.; Darabi, H.; Haghizadeh, A.; Haghighi, A.T.; Nalivan, O.A.; Bui, D.T. The effect of sample size on different machine learning models for groundwater potential mapping in mountain bedrock aquifers. *Catena* **2020**, *187*, 104421. [\[CrossRef\]](#)
22. Velásquez, P.L.C.; Zuluaga, A.B.; Hernández, M.B.; Estrada, A.L. *Plan de Manejo y Ordenamiento Ambiental del Complejo Cenagoso Del Bajo Sinú*; Corporación Autónoma Regional de Los Valles Del Sinú y Del San Jorge: Medellín, Colombia; Universidad Nacional de Colombia: Bogotá, Colombia, 2008.
23. Vargas, R.D.S.; Caro, M.A.T.; Doria, D.D.F. Conflictos por el agua y resiliencia comunitaria en el Bajo Sinú: Evidencias de disputas y alternatividad. *Utopía y Prax. Latinoam. Rev. Int. Filos. Iberoam. y Teoría Soc.* **2020**, *25*, 105–124.
24. IDEAM. *Tiempo y Clima, Reportes. Boletín Climatológico*; Instituto de Hidrología, Meteorología y Estudios Ambientales: Bogotá, Colombia, 2018.
25. IDEAM; INVERMAR. Protocolo de Monitoreo Del Agua. 2017. Available online: https://www.ideam.gov.co/nuestra-entidad/hidrologia/protocolo-de-monitoreo-y-seguimiento-del-agua-PROTOCOLO_MONITOREO_AGUA_IDEAM.pdf (accessed on 5 July 2024).
26. Verma, D.K.; Satyaveer, N.K.M.; Kumar, P.; Jayaswa, R. Important water quality parameters in aquaculture: An overview. *Agric. Environ.* **2022**, *3*, 24–29.
27. Benesty, J.; Chen, J.; Huang, Y.; Cohen, I. Pearson correlation coefficient. In *Noise Reduction in Speech Processing*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 1–4. [\[CrossRef\]](#)
28. O'Brien, R.M. A caution regarding rules of thumb for variance inflation factors. *Qual. Quant.* **2007**, *41*, 673–690. [\[CrossRef\]](#)
29. Miles, J. Tolerance and variance inflation factor. In *Encyclopedia of Statistics in Behavioral Science*; John Wiley & Sons, Ltd.: Hoboken, NJ, USA, 2005. [\[CrossRef\]](#)
30. Li, J.; Heap, A.D. *A Review of Spatial Interpolation Methods for Environmental Scientists*; Geoscience Australia: Canberra, Australia, 2008.
31. Demir, E.; Saatçioğlu, Ö.; İmrol, F. Uluslararası dergilerde yayımlanan eğitim araştırmalarının normallik varsayımları açısından incelenmesi. *Curr. Res. Educ.* **2016**, *2*, 130–148.
32. Kumari, M.; Sarma, K.; Sharma, R. Using Moran's I and GIS to Study the Spatial Pattern of Land Surface Temperature in Relation to Land Use/Cover around a Thermal Power Plant in Singrauli District, Madhya Pradesh, India. *Remote Sens. Appl. Soc. Environ.* **2019**, *15*, 100239. [\[CrossRef\]](#)
33. Yang, H.; Kong, J.; Hu, H.; Du, Y.; Gao, M.; Chen, F. A Review of Remote Sensing for Water Quality Retrieval: Progress and Challenges. *Remote Sens.* **2022**, *14*, 1770. [\[CrossRef\]](#)
34. De Smith, M.J.; Goodchild, M.F.; Longley, P.A. *Geospatial Analysis: A Comprehensive Guide to Principles, Techniques and Software Tools*; Troubador Publishing Ltd.: Leicestershire, UK, 2007. Available online: <https://books.google.es/books?hl=es&lr=&id=SULMdT8qPwEC&oi=fnd&pg=PA7&dq=Geospatial+analysis:+a+comprehensive+guide+to+principles,+techniques+and+software+tools&ots=fovz7LpH7T&sig=ON3vZFl6f3JDuz59vGBNHLwEimY> (accessed on 6 October 2025).
35. Childs, C. Interpolating surfaces in ArcGIS spatial analyst. In *ArcUser*; ESRI: Redlands, CA, USA, 2004; Volume 3235, pp. 32–35.
36. Tan, Q.; Xu, X. Comparative Analysis of Spatial Interpolation Methods: An Experimental Study. *Sens. Transducers* **2014**, *165*, 155.
37. Burgess, T.M.; Webster, R. Optimal Interpolation and Isarithmic Mapping of Soil Properties: I The Semi-variogram and Punctual Kriging. *Eur. J. Soil Sci.* **2019**, *70*, 11–19. [\[CrossRef\]](#)
38. Krivoruchko, K. Empirical bayesian kriging. *ArcUser Fall* **2012**, *6*, 1145.
39. Fox, J. *Applied Regression Analysis and Generalized Linear Models*; Sage Publications: Thousand Oaks, CA, USA, 2015.
40. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [\[CrossRef\]](#)

41. Biau, G.; Scornet, E. A random forest guided tour. *Test* **2016**, *25*, 197–227. [\[CrossRef\]](#)
42. Schonlau, M.; Zou, R.Y. The random forest algorithm for statistical learning. *Stata J.* **2020**, *20*, 3–29. [\[CrossRef\]](#)
43. Halder, R.K.; Uddin, M.N.; Uddin, M.A.; Aryal, S.; Khraisat, A. Enhancing K-nearest neighbor algorithm: A comprehensive review and performance analysis of modifications. *J. Big Data* **2024**, *11*, 113. [\[CrossRef\]](#)
44. Natekin, A.; Knoll, A. Gradient boosting machines, a tutorial. *Front. Neurobotics* **2013**, *7*, 21. [\[CrossRef\]](#) [\[PubMed\]](#)
45. Johnson, R.; Zhang, T. Learning nonlinear functions using regularized greedy forest. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *36*, 942–954. [\[CrossRef\]](#)
46. Riedmiller, M.; Lernen, A. *Multi Layer Perceptrón; Machine Learning Lab Special Lecture*; University of Freiburg: Freiburg im Breisgau, Germany, 2014; Volume 24, pp. 11–60.
47. Jiang, Q.; Zhu, L.; Shu, C.; Sekar, V. An efficient multilayer RBF neural network and its application to regression problems. *Neural Comput. Appl.* **2022**, *34*, 4133–4150. [\[CrossRef\]](#)
48. Renaud, O.; Victoria-Feser, M.-P. A robust coefficient of determination for regression. *J. Stat. Plan. Inference* **2010**, *140*, 1852–1862. [\[CrossRef\]](#)
49. Hodson, T.O. Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not. *Geosci. Model Dev. Discuss.* **2022**, *15*, 5481–5487. [\[CrossRef\]](#)
50. Chai, T.; Draxler, R.R. Root mean square error (RMSE) or mean absolute error (MAE). *Geosci. Model Dev. Discuss.* **2014**, *7*, 1525–1534.
51. Chicco, D.; Warrens, M.J.; Jurman, G. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Comput. Sci.* **2021**, *7*, e623. [\[CrossRef\]](#)
52. Li, J. Assessing spatial predictive models in the environmental sciences: Accuracy measures, data variation and variance explained. *Environ. Model. Softw.* **2016**, *80*, 1–8. [\[CrossRef\]](#)
53. Fouedjio, F.; Talebi, H. Geostatistical semi-supervised learning for spatial prediction. *Artif. Intell. Geosci.* **2022**, *3*, 162–178. [\[CrossRef\]](#)
54. Huang, C.; Shibuya, A. High Accuracy Geochemical Map Generation Method by a Spatial Autocorrelation-Based Mixture Interpolation Using Remote Sensing Data. *Remote Sens.* **2020**, *12*, 1991. [\[CrossRef\]](#)
55. Cressie, N. *Statistics for Spatial Data*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
56. Wise, S. Cross-validation as a means of investigating DEM interpolation error. *Comput. Geosci.* **2011**, *37*, 978–991. [\[CrossRef\]](#)
57. Ejigu, M.T. Overview of water quality modeling. *Cogent Eng.* **2021**, *8*, 1891711. [\[CrossRef\]](#)
58. Lv, J.Q.; Yin, W.X.; Xu, J.M.; Cheng, H.Y.; Li, Z.L.; Yang, J.X.; Wang, A.-J.; Wang, H.C. Augmented machine learning for sewage quality assessment with limited data. *Environ. Sci. Ecotechnol.* **2025**, *23*, 100512. [\[CrossRef\]](#)
59. Masud, M.M.; Shamem, A.S.M.; Saif, A.N.M.; Bari, M.F.; Mostafa, R. The role of artificial intelligence in sustainable water management in Asia: A systematic literature review with bibliographic network visualization. *Int. J. Energy Water Resour.* **2025**, *9*, 247–265. [\[CrossRef\]](#)
60. Staehr, S.U.; Van der Zande, D.; Staehr, P.A.; Markager, S. Suitability of multisensory satellites for long-term chlorophyll assessment in coastal waters: A case study in optically-complex waters of the temperate region. *Ecol. Indic.* **2022**, *134*, 108479. [\[CrossRef\]](#)
61. Bouasria, A.; Bouslihim, Y.; Gupta, S.; Taghizadeh-Mehrjardi, R.; Hengl, T. Predictive performance of machine learning model with varying sampling designs, sample sizes, and spatial extents. *Ecol. Inform.* **2023**, *78*, 102294. [\[CrossRef\]](#)
62. Rajput, D.; Wang, W.J.; Chen, C.C. Evaluation of a decided sample size in machine learning applications. *BMC Bioinform.* **2023**, *24*, 48. [\[CrossRef\]](#)
63. Ramezan, C.A.; Warner, T.A.; Maxwell, A.E.; Price, B.S. Effects of training set size on supervised machine-learning land-cover classification of large-area high-resolution remotely sensed data. *Remote Sens.* **2021**, *13*, 368. [\[CrossRef\]](#)
64. Schmidinger, J.; Schröter, I.; Bönecke, E.; Gebbers, R.; Ruehlmann, J.; Kramer, E.; Mulder, V.L.; Heuvelink, G.B.M.; Vogel, S. Effect of training sample size, sampling design and prediction model on soil mapping with proximal sensing data for precision liming. *Precis. Agric.* **2024**, *25*, 1529–1555. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.