hydrology

MDPI

Article

Machine Learning Analysis of Hydrological and Hydrochemical Data from the Abelar Pilot Basin in Abegondo (Coruña, Spain)

Javier Samper-Pilar [1], Javier Samper-Calvete [1,*], Alba Mon [1], Bruno Pisani [1] and Antonio Paz-González [2]

[1] Interdisciplinary Center of Chemistry and Biology (CICA), Civil Engineering, Campus de Elviña, University of A Coruña, 15071 A Coruña, Spain; javier.samper@fuac.udc.es (J.S.-P.); alba.mon@udc.es (A.M.); bruno.pisani@udc.es (B.P.)

[2] Interdisciplinary Center of Chemistry and Biology (CICA), Facultad de Ciencias, University of A Coruña, 15071 A Coruña, Spain; antonio.paz.gonzalez@udc.es

* Correspondence: j.samper@udc.es

Abstract: The Abelar pilot basin in Coruña (northwestern Spain) has been monitored for hydrological and hydrochemical data to assess the effects of eucalyptus plantation and manure applications on water resources, water quality, and nitrate contamination. Here, we report the machine learning analysis of hydrological and hydrochemical data from the Abelar basin. K-means cluster analysis (CA) is used to relate nitrate concentrations at the outlet of the basin with daily interflows and groundwater flows calculated with a hydrological balance. CA identifies three linearly separable clusters. Times series Gaussian process regression (TS-GPR) is employed to predict surface water nitrate concentration by incorporating hydrological variables as additional input parameters using a time series shifting. TS-GPR allows modelling nitrate concentrations based on shifted interflows and groundwater flows and chemical concentrations with R^2 = 0.82 and 0.80 for training and testing, respectively. Groundwater flow from five days prior to the current date, Q_{g5}, is the most important input parameter of the TS-GPR model. Interaction effects between the variables are found. TS-GPR validation with recent data provides results consistent with those of testing (R^2 = 0.85). Model inspection by permutation feature importance and partial dependence plots shows interactions between Q_{g5} and Cl, and between Ca and Mg.

Keywords: K-means clustering; Gaussian process regression; time series analysis; Abelar pilot basin; nitrate concentration; applied machine learning; hydrological data; hydrochemical data

check for updates

Academic Editors: Osman Fetoshi, Gabriel-Ionut Plavan and Bojan Đurin

Received: 31 January 2025
Revised: 3 March 2025
Accepted: 4 March 2025
Published: 6 March 2025

Citation: Samper-Pilar, J.; Samper-Calvete, J.; Mon, A.; Pisani, B.; Paz-González, A. Machine Learning Analysis of Hydrological and Hydrochemical Data from the Abelar Pilot Basin in Abegondo (Coruña, Spain). *Hydrology* **2025**, *12*, 49. https://doi.org/10.3390/hydrology12030049

Copyright: © 2025 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) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In the last two decades, machine learning (ML) has emerged as a transformative tool for analysing complex and high-dimensional hydrological and hydrochemical datasets, offering novel insights and predictive capabilities. Datasets are often high-dimensional, with temporal and spatial variability posing significant challenges. Common ML techniques for analysing river streamflow and water quality data include (1) supervised learning with algorithms such as artificial neural networks (ANNs), random forests, support vector machines, and gradient boosted trees are widely used for classification and regression tasks; (2) deep learning: recurrent neural networks (RNNs) and long short-term memory (LSTM) networks are effective for time-series prediction, while convolutional neural networks (CNNs) are applied to spatial pattern recognition; (3) unsupervised learning such as clustering methods, dimensionality reduction techniques, principal component analysis (PCA), and T-distributed stochastic neighbour embedding (t-SNE) assist in identifying

patterns and anomalies in water quality data; and (4) hybrid models, which integrate ML with traditional hydrological models such as SWAT and MODFLOW, have improved the prediction accuracy and the interpretation of model results. A notable advancement in this category is physics-informed neural networks (PINNs), which embed governing physical equations within deep learning frameworks to jointly estimate parameters and states, making them particularly effective for subsurface transport and sparse-data scenarios.

ANNs have been used to estimate nitrate concentrations in groundwater based on in situ field data, with results demonstrating that incorporating land use variables significantly improved model performance in terms of root-mean-square error (RMSE) offering a promising tool for managing groundwater contamination in the Asopos River Basin and Kopaidian Plain, Greece [1]. Random forests were compared with linear modelling to estimate the concentration of nutrients including nitrate in a rural catchment, using commonly measured in situ variables as surrogates, leading to a reduction in RMSE of up to 60.1% [2]. Random forests outperformed other ML models in predicting total nitrogen concentrations in inland rivers with 92.94% accuracy and were also used to model salinity using remote sensing data. Support vector machines and gradient boosting effectively predicted water quality indicators, such as turbidity and nutrient levels, across diverse watersheds [3]. Random forest regression was found to provide accurate predictive models for nitrate pollution in groundwater, identifying key predictors of pollution, and outperforming logistic regression in generating vulnerability maps for water resources management [4].

LTSM networks effectively modelled complex, nonlinear, and temporal dependencies in hydrological data, outperforming traditional models in streamflow prediction, particularly under high variability, and offering improved predictive capabilities and potential for real-time hydrological forecasting [5]. A LSTM model was also used to predict river dissolved oxygen dynamics, effectively capturing its relationship with water temperature across 236 U.S. watersheds and highlighting the potential of deep learning models for predicting river water quality at large scales [6]. Standalone LSTM and CNN models, along with a coupled CNN-LSTM model, were developed to predict water quality variables—dissolved oxygen (DO) and chlorophyll-a (Chl-a)—in Small Prespa Lake, Greece, using time-series data of physicochemical variables. The coupled CNN-LSTM model outperforms both standalone and traditional ML models, demonstrating its ability to capture both low and high levels of water quality variables. LSTM performs the best for DO prediction [7].

Hierarchical cluster analysis (HCA) was used to analyse hydrochemical datasets, comparing different linkage methods to optimize classification for complex hydrological systems [8]. HCA and Mahalanobis distance metrics were employed to monitor human-induced changes in the Jinjiang River's hydrochemistry, highlighting pollution impacts from agriculture and mining [9]. t-SNE was introduced as a graphic approach to assist HCA in groundwater geochemistry, outperforming PCA in determining the number of clusters and delineating spatial geochemical zones. However, t-SNE reliance on hyperparameter tuning limits its standalone use [10]. A hybrid model approach combining LSTM networks with the SWAT model was proposed to improve streamflow prediction, demonstrating the benefits of integrating ML with traditional hydrological models for more accurate and interpretable predictions [11]. Previous studies have focused on the use of K-means clustering to provide information on processes affecting aquifers [12,13]. Recent studies have demonstrated the effectiveness of unsupervised clustering in reservoir water quality assessment, reinforcing the potential of classification-based methods for structuring hydrochemical data [14]. A multiphysics-informed deep neural network approach based on PINNs was proposed for estimating space-dependent hydraulic conductivity, hydraulic head, and concentration fields from sparse measurements by jointly training deep neu-

ral networks with governing equation residuals, demonstrating superior accuracy over standard data-driven methods, particularly in sparse data scenarios [15].

Regression models have been used very scarcely in the field of groundwater quality, and mainly through linear regression and multiple linear regression methods as the benchmark for the lowest acceptable accuracy [16]. Gaussian process regression (GPR) was found to be an effective tool in predicting nitrate concentration in groundwater based only on water quality chemical input parameters, outperforming popular decision tree algorithms [17]. Similarly, GPR consistently yielded good performance in predicting surface water nitrate concentration across various watersheds with differing land use practices, including agriculture and forest land use, using streamflow as the only input parameter [18]. Recent research highlighted the effectiveness of GPR in hydrological modelling, successfully applying it to forecast nonstationary hydrological datasets and capture complex relationships in hydrological systems [19]. A multi-scale Gaussian process framework was recently developed to enhance efficiency, optimize GPR for large-scale hydrological applications, and address the scalability challenges of traditional GPR models with large datasets [20].

The water supply of the population in Atlantic European regions living in dispersed rural areas is managed privately through autonomous solutions, such as domestic wells or neighbourhood water systems that capture springs. There is growing concern about microbiological and nitrate contamination of groundwater. This contamination is caused by the improper management of organic fertilizers in agriculture, slurry discharges from farms, and land use planning that does not consider water sources [21,22]. Implementing European regulations on drinking water (Directive 98/83/EC) and water protection and management (Directive 2000/60/EC) in these communities poses a significant challenge for water administration, as the feasibility of centralized infrastructure is limited by the enormous investments required and the high maintenance costs. More than 50% of the population of Abegondo (A Coruña, Spain) relies on autonomous water systems [22]. The main issues for the sustainability of these systems include (1) insufficient supply reliability; (2) deterioration of water quality; and (3) deficiencies in the governance of groundwater use and waste management [23]. Samper et al. [21] presented the results of a hydrogeological study that analysed the vulnerability of groundwater to contamination and the protection of water sources in rural areas of Abegondo (A Coruña). Subsequently, Samper et al. [23] presented an analysis of the sustainability of autonomous groundwater supply systems in the rural areas of Abegondo, which are located in fractured hard rocks. There is a concern about the chemical and microbiological quality of groundwater supply in rural areas due to the lack of proper control of the water quality by well owners. Agricultural contamination is often a pressure on the groundwater chemical status where nitrate is the main concern for groundwater quality.

Hydrological research and water resources assessment were performed in the Abelar pilot watershed in Abegondo by Soto et al. [24] and Rodríguez-Suárez et al. [25] (Figure 1). This small watershed has a surface area of about 10.7 ha. It was equipped with two meteorological stations, two piezometers, and a streamflow gauging station [20,21]. The results show that fast-growing trees with high water consumption can reduce the availability of water resources, especially in the summer months and other dry periods, which can be manifested in the decrease in piezometric levels and the eventual drying up of springs and riverbeds. A hydrometeorological water balance model of the Abelar site was presented by Samper et al. [23] to quantify the water resources in this basin planted with eucalyptus. The hydrological model of the Abelar basin was calibrated with piezometric and streamflow data [23].

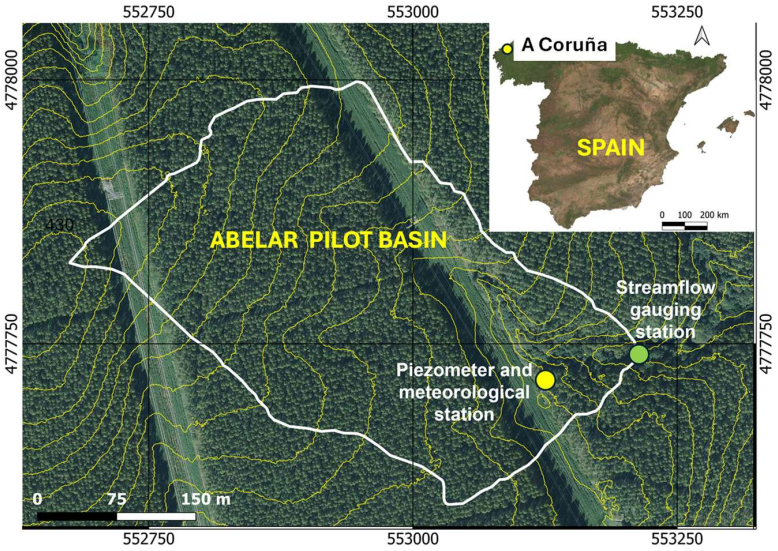


Figure 1. Location of the Abelar pilot basin south of A Coruña city in northwestern Spain. The coordinate system is ETRS89/UTM zone 29N (EPSG 25,829). Ground elevation contour lines are spaced 2 m. The elevation of the gauging station is 394 m a.s.l.

Here, we present the machine learning analysis of hydrological and hydrochemical data from the Abelar basin. K-means clustering is used to identify groups for nitrate concentration and hydrological flows (interflow and groundwater flow). Our work also further explores the underrepresented regression model of GPR, extending its scope to predict surface water nitrate concentration by incorporating hydrological variables as additional input parameters using a time series shifting approach. Specifically, the methods of cluster analysis (CA) and GPR for time series (TS-GPR) are applied to group and predict surface water nitrate concentration. The paper starts with the description of the Abelar pilot basin, the available data, and the ML methods used. Then, model results are presented for the ML methods. Afterward, a discussion of the main findings is presented. The paper ends with the main conclusions and future work suggestions to improve the individual performance of the ML methods as well as their potential coupling.

2. Materials and Methods

2.1. Abelar Pilot Basin and Available Data

The Abelar pilot basin has been monitored for hydrological and hydrochemical data during the last few decades. These data have been used to assess the effects of eucalyptus plantation on water resources, soil properties, water quality, and nitrate contamination due to excessive manure applications. The Abelar basin is located at the catchment of the Tambre river basin (Figure 1). The average elevation of the basin is 413 m above sea level. The climate is humid and oceanic with relatively abundant rainfall. The mean annual precipitation in the period 1997–2016 is equal to 1577 mm. Temperatures are mild and show little oscillation between maximum and minimum (16.7 °C in summer and 8.1 °C in winter). November and December are the rainiest months. The climate of the study area is of the Csb type according to the Köppen–Geiger classification [26]. The basin is located on schists (metapsamites and metapelites) of the Betanzos Unit. The weathering and alteration of shale gives rise to a surface layer of variable thickness between 1 and 2 m, in some cases reaching 5 m [22]. The vegetation consists of *Eucalyptus globulus* and low shrubs. Groundwater recharge occurs mainly due to the infiltration of rainwater. The discharge of diffuse or localized groundwater flow occurs in the valleys and low areas of the basin. The basin of the farm has been characterized continuously since the hydrological year 1997/98

in which the plantation was carried out and an automatic and a manual weather station, an automatic streamflow gauging station, and two piezometers were installed.

The characterization of the main creek in the Abelar pilot basin at the in situ automatic streamflow gauging station has produced a wealth of hydrochemical data since its installation. The focus of this study is on analysing the combination of such surface water hydrochemical data with meteorological data and hydrological model outputs such as daily groundwater recharge rates, interflows, and groundwater flows [22]. The Abelar dataset contains data on hydrological and hydrochemical properties of the Abelar basin in Abegondo, La Coruña (Spain). The specific attributes of the Abelar dataset given are listed in Table 1. The dataset contains 9679 entries of daily meteorological, hydrological, and hydrochemical data corresponding to the period from 1 October 1997 to 31 March 2024. While meteorological and hydrological data are available every day, the hydrochemical data are available at some specific dates corresponding to the water sampling dates from 26 January 2007 to 30 July 2024. The number of data is 403 for all species except for Fe (401 observations), Cu (393 observations), Zn (397 observations), Al (398 observations), and Vn (394 observations). Subsequently, additional hydrochemical data up to 30 August 2024 were obtained and the simulated hydrological data extended accordingly. Simulations of hydrological data were performed with the code VISUAL BALAN v2.0 (e.g., [27–29]), which is based on a semi-distributed model that performs daily water balances in the soil, the underlying unsaturated zone, and the aquifer. VISUAL BALAN evolved from earlier versions of the code, which had the generic name of BALAN [27,30]. The hydrological components are evaluated daily in a sequential manner. The daily values of precipitation and streamflow components are shown in Figure 2.

Table 1. Attributes of the Abelar dataset.

Name of Attribute	Description	Units
Date	Date of the observation	-
P	Precipitation	mm
R	Recharge	mm
Q_s	Surface runoff	mm
Q_i	Interflow/subsurface flow	mm
Q_g	Groundwater flow	mm
Q_t	Total flow	mm
K	Potassium concentration	mg/L
Na	Sodium concentration	mg/L
Ca	Calcium concentration	mg/L
Mg	Magnesium concentration	mg/L
Fe	Iron concentration	µg/L
Mn	Manganese concentration	µg/L
Cu	Copper concentration	µg/L
Zn	Zinc concentration	µg/L
Al	Aluminium concentration	µg/L
Vn	Vanadium concentration	µg/L
Si	Silicon concentration	mg/L
Cl	Chloride concentration	mg/L
SO_4^{2-}	Sulphate concentration	mg/L
NO_3^-	Nitrate concentration	mg/L

Figure 3 shows the temporal evolution of concentration data. These data are grouped based on the following three conditions: (1) predominant interflow with interflow being greater than 75% of the total flow; (2) predominant groundwater flow with groundwater being more than 80% of the total flow; and (3) interflow fraction ranging from 20% to 75%. Nitrate, chloride, calcium, magnesium, and potassium concentrations at the basin outlet were collected from January 2007 to December 2015. The second sampling campaign

started in the spring of 2023 and includes data until November 2024. Nitrate concentrations range from 10 to 40 mg/L, with two peaks of 150 mg/L. Nitrate concentrations fluctuate in response to hydrological variations in the basin. Higher nitrate concentrations are associated with an increase in interflow, while lower concentrations correspond to larger groundwater flows. The lowest nitrate concentrations, typically observed during the dry (low flow) season, are considered more representative of groundwater concentrations. Chloride, calcium, magnesium, and potassium concentrations exhibit trends like nitrate, although with less pronounced variations.

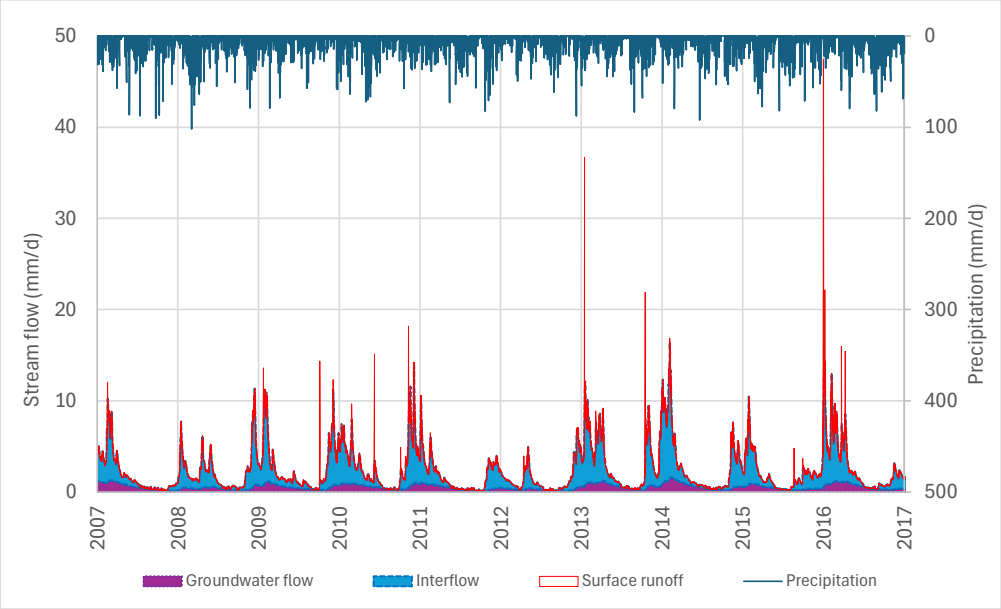


Figure 2. Stacked area chart displaying daily values of precipitation and streamflow components.

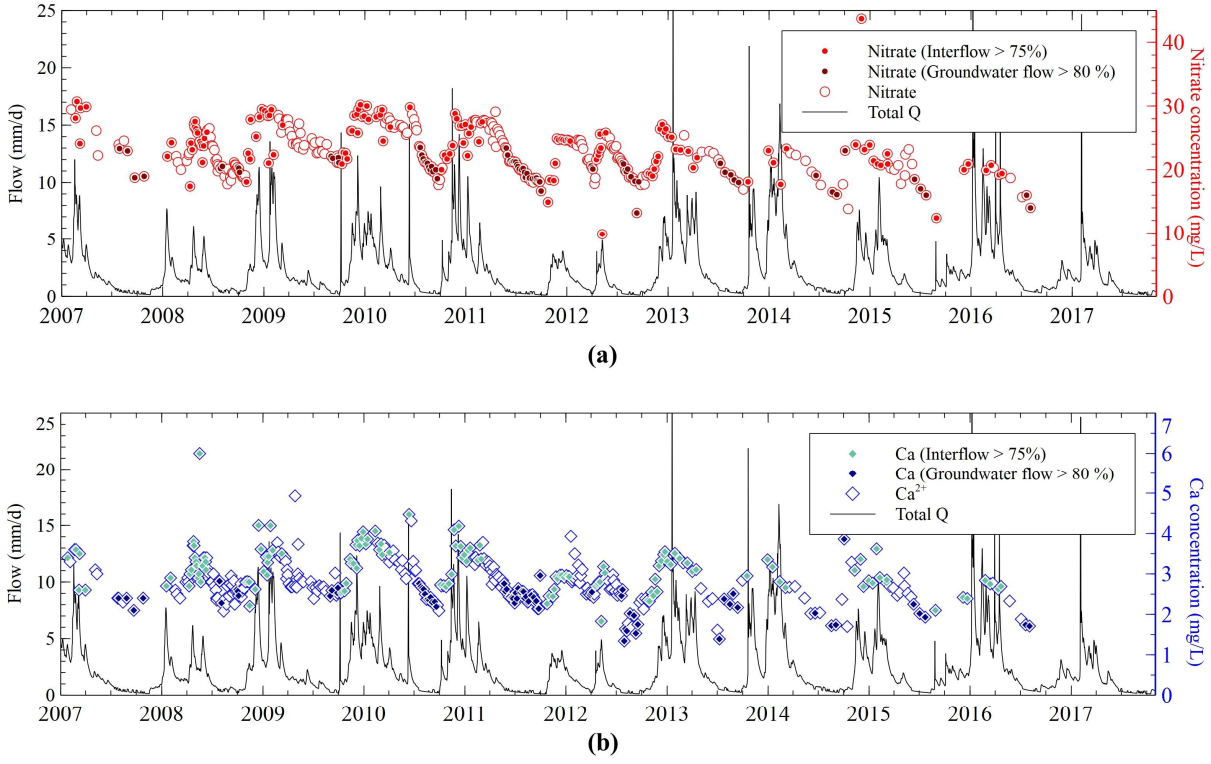


Figure 3. Time evolution of daily streamflows calculated with the hydrological water balance model (lines) and measured chemical data of chloride and nitrate (**a**) and Ca, Mg, and K data (**b**).

2.2. ML Methods

Cluster analysis (CA) and (2) time series Gaussian process regression (TS-GPR) were used to analyse the selected hydrochemical target variables. Data preprocessing was performed for both methods prior to modelling, as required by ML techniques.

2.2.1. Data Preprocessing

Data preparation was conducted by following two slightly different processes for CA and TS-GPR ahead of modelling.

The data preprocessing before the CA included data exploration through histograms, heatmaps, and pair plots of the data to look for correlation among variables. Feature engineering was conducted prior to the K-means modelling, including the log-transformation of all input variables. This transformation involved applying the natural logarithm to one plus the variable's value to account for values equal to or near zero. Log-transforming the input variables helped address their skewed distributions. Notably, the target variable (nitrate) was not subjected to log-transformation.

The data preparation for TS-GPR began with resampling the data to a daily frequency to account for the irregular frequency of sampling of chemical data. Days with several chemical samplings were resampled by taking the mean value of the observations for each date. Anomalous values were identified and removed from the data using the quantile-based outlier detection (QBOD) method. QBOD is a technique that utilizes statistical quantiles to detect outliers by comparing each data point to a range defined by quantiles. It extends the interquartile range (IQR) analysis introduced by Tukey [31], incorporating flexible quantile-based thresholds to make it more robust to skewed distributions. The number of anomalous values was defined by setting lower and upper quantile bounds. Preliminary analyses determined that using quantiles of 0.25 and 0.75 provided the best balance between retaining informative data and excluding extreme anomalous data that could skew the model results. Four anomalous samples were identified and removed from nitrate data, corresponding to the top four values in the distribution.

2.2.2. Cluster Analysis

Cluster analysis involves classifying an integrated hydrological system (defined as the combination of hydrological and hydrochemical data) into sub-hydrological systems based on the ratio of interflow to total flow (Q_1), the ratio of groundwater flow to total flow (Q_2), and nitrate concentration. The clustering K-means algorithm implemented in the Scikit-learn package 1.5.2 [32] of the Python programming language (version 3.12.3) was used in this approach. The unsupervised K-means algorithm [33] is an iterative process in which similar observations are grouped together. The algorithm starts by taking two random points known as centroids and continues by calculating the distance of each observation to the centroid and assigning each cluster to the nearest centroid. After the first iteration, every point belongs to a cluster. Next, the number of centroids increases by one, and the centroid for each cluster is recalculated as the points with the average distance to all points in a given cluster. The process is repeated k-times until no observation is assigned to another cluster. The algorithm converges when clusters do not move anymore.

CA for the groups was carried out by fitting 20 K-means models on the data with values of the cluster number ranging from 1 to 20 and storing for each model the number of clusters and the standard inertia value. The model fitting was performed using the KMeans class from the *sklearn.cluster* module in Scikit-learn. The random state parameter was set equal to 42 to ensure reproducibility, while the default parameters for maximum number of iterations (max_iter) and convergence tolerance (tol) were set equal to 300 and 10^{-4}, respectively. The analysis was conducted on a system running on Windows 11

using NumPy 1.26.4 and Pandas 1.5.3. The resulting dataset after data preprocessing consisted of 388 samples (after removing four anomalous values), all of which were used for training. The input variables or features for the K-means models were Q_1, Q_2, and nitrate concentration. The optimal number of clusters was then visually determined based on the plot of inertia versus the number of clusters (elbow method). The final clusters were both visualized with pair plots and statistically analysed. The visualization was conducted with the following Python libraries: Matplotlib (version 3.9.2) and Seaborn (version 0.13.2). The statistical computations were performed with Pandas. The methodology for the CA model applied in this study is outlined in Figure 4.

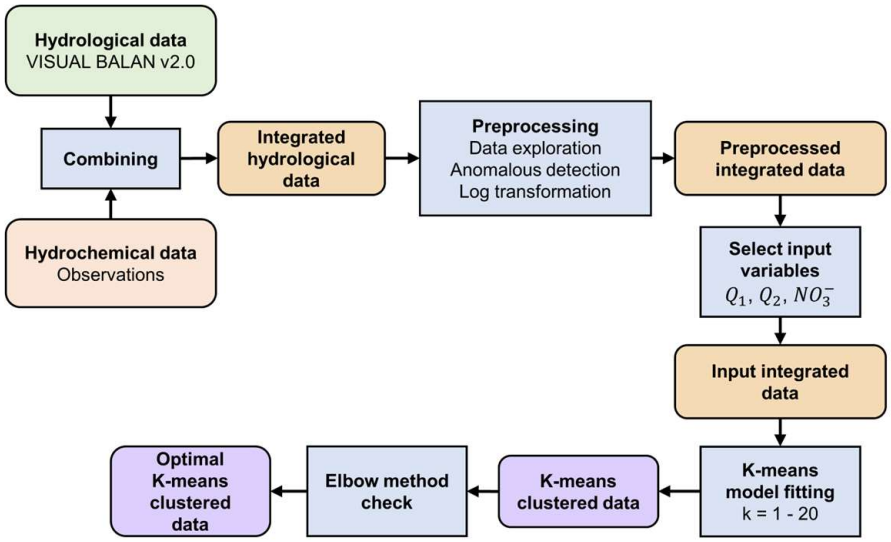


Figure 4. Flowchart of the methodology of CA modelling in this study.

The main advantage of the K-means algorithm is that it is easy to compute. A disadvantage is that this algorithm is sensitive to the choice of the initial points, so different initial configurations may yield different results. To resolve this, a more intelligent initialization method for K-means clusters, known as K-means ++, helps avoid falling into local optima. This is the default implementation of the K-means in Scikit-learn that was used here.

2.2.3. Times Series Gaussian Process

Gaussian process (GP) is a nonparametric supervised learning method commonly used to solve regression and probabilistic classification problems [34]. GP is an exact interpolator for noise-free data and regular covariance functions (kernels). GP predictions are probabilistic and therefore GP allows for computing approximate confidence intervals. The Gaussian process algorithm aims at estimating a real-valued function, $f(\boldsymbol{x})$, defined on a compact subset $A \subset \mathbb{R}^d$ using a finite set of data values, $y_i = f(\boldsymbol{x}_i)$, corresponding to locations $\boldsymbol{x}_i$ with $i = 1, 2, \ldots t$, where t is the number of training data points and d is the dimensionality of function f. It is assumed that f is a GP having a covariance function $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ which quantifies the covariance of f at locations $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. GP provides unbiased and minimum variance estimates of f conditioned on the training dataset of t observations. The posterior distribution of the GP has a mean μ_t and a covariance function $k_t(\boldsymbol{u}, \boldsymbol{v})$ which are given by:

$$\mu_t(\boldsymbol{u}) = k_t(\boldsymbol{u})^{\mathrm{T}} K_t^{-1} \boldsymbol{y}_t \quad (1)$$

$$k_t(\boldsymbol{u}, \boldsymbol{v}) = k(\boldsymbol{u}, \boldsymbol{v}) - k_t(\boldsymbol{u})^{\mathrm{T}} K_t^{-1} k_t(\boldsymbol{v}) \quad (2)$$

where $\boldsymbol{u}$ and $\boldsymbol{v}$ are column vectors of the locations of the t query or testing points; $\boldsymbol{u} = [u_1, u_2, \ldots, u_d]^T$ and $\boldsymbol{v} = [v_1, v_2, \ldots, v_d]^T$ in the input domain $A \subset \mathbb{R}^d$ where

the GP algorithm predicts the values of f or computes the covariance matrix $\boldsymbol{K}_t$, which has entries $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ with $i,j = 1, 2, \ldots t$; $k_t(\boldsymbol{u}, \boldsymbol{v})$ is the covariance of the estimates of f at locations $\boldsymbol{u}$ and $\boldsymbol{v}$ and $k_t(\boldsymbol{u})$ and $\boldsymbol{y}_t$ are t-dimensional column vectors given by $k_t(\boldsymbol{u}) = [k(\boldsymbol{x}_1, \boldsymbol{u}) \ldots k(\boldsymbol{x}_t, \boldsymbol{u})]^T$ and $\boldsymbol{y}_t = [y_1, y_2, \ldots, y_t]^T$, respectively; and T denotes the vector/matrix transpose. For further details on GP, we refer the reader to Rasmussen and Williams [34], which thoroughly presents the properties of GP.

A kernel is a crucial ingredient of GP that determines the shape of the prior and posterior distributions of the GP. The covariance $k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ of a stationary GP depends only on the Euclidian distance d between the two points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. Here, we used the Matérn stationary kernel, which is a generalization of the radial basis function kernel [35] and is given by:

$$k(d) = \frac{1}{\Gamma(v)2^{v-1}} \left(\frac{\sqrt{2v}}{l} d \right)^v K_v \left(\frac{\sqrt{2v}}{l} d \right) \quad (3)$$

where $K_v(\cdot)$ is a modified Bessel function, $\Gamma(\cdot)$ is the gamma function, l (>0) is a length-scale parameter, and v is a parameter controlling the smoothness of the kernel covariance function.

In the system considered here, f models a physical quantity, e.g., species concentration, that is assumed to be smooth and k is either the squared exponential covariance function or the Matérn covariance function, both commonly used in this case.

The method of TS-GPR implements a Gaussian process regressor approach that models temporal dependencies by defining a kernel over sequential data, enabling complex pattern recognition and uncertainty quantification over time. Sequential data for time series were generated through data shifting, a feature engineering technique that creates lagged features from the available data to model temporal dependencies. Hence, once the target variable was selected, the hydrological variables of the dataset were shifted, using time lags up to 10 days, while the hydrochemical variables were kept fixed at the same date as that of the observations of the target variable. Using a time lag of 1 day is equivalent to keeping the parameters fixed. As a result of data shifting, a dataset of shape $m \times n$ was generated, where m is the number of days on which both hydrological and hydrochemical data are available, and n represents the number of total variables after shifting, including the target variable.

The calculations of TS-GPR were carried out using the GaussianProcessRegressor class from the *sklearn.gaussian*_process module in Scikit-learn (version 1.5.2) using the Python programming language (version 3.11.9). A Matérn kernel with v = 0.5 was employed to capture the temporal dependencies in the data, with initial length scales set to an array of ones corresponding to the number of input features (N). The length scales were constrained within the bounds (10^{-3}, 10^2) to allow for adaptive optimization. The Gaussian process model was fitted with an additive noise term, alpha, equal to 10^{-7} to enhance numerical stability and was optimized using 10 restarts of the built-in optimizer and setting *n_restarts_optimizer* = 10. The target variable was standardized while training by enabling *normalize_y* = *True*. The analysis was conducted on a system running on Windows 11 using NumPy 1.26.4 and Pandas 2.2.3.

Only data up to the year 2016 were used for training and testing the model. Observations from that year onward were used for validation purposes. Trial runs suggested that optimal results were consistently achieved by transforming the input variables with a scaler and the target variable with another. Each of the two scalers worked by subtracting the mean and dividing by the standard deviation of each variable's samples. Of the data, 80% were used for training and 20% for testing. Trial runs confirmed that the best results were consistently obtained by initializing the bounds of the length scales (10^{-3}, 10^2). Similarly, the choice of the Matérn kernel with v = 0.5, which controls the smoothness of the learned

function, was validated as the most appropriate for the data. Some trials were made to automatically select the hydrochemical variables, but no convincing results were obtained. Test runs were also conducted to iteratively fit successive TS-GPR models and eliminate the less relevant features, but the results were not compelling.

The TS-GPR modelling was performed in four steps with increasing number of input variables and time lag days for shifting, with nitrate concentration as the target variable. The steps include the following models: (1) a baseline model that uses the classical method of predicting nitrate concentration based solely on total flow (Q_t) without shifting; (2) a model with the best correlated hydrological variables R, Q_i (interflow), and Q_g (groundwater flow) without shifting; (3) a model with hydrological variables R, Q_i, and Q_g with shifting; and (4) the most complex model with hydrological variables R, Q_i, and Q_g with shifting and the best correlated hydrochemical variables. To confirm the reliability, robustness, and confidence in each model's performance estimates, repeated k-fold cross validation was performed on the training set at each step of the modelling process to obtain a more thorough evaluation across multiple train–test splits. Thus, the training results at each step are based on k-fold cross-validation of the training set with 10 folds and 10 repeats.

The TS-GPR modelling results were visualized with scatter plots and time series plots using the Python libraries Matplotlib (version 3.9.2) and Seaborn (version 0.13.2).

2.2.4. Model Inspection Methods

A critical aspect of building trustworthy and interpretable predictive models is the ability to inspect and understand how each variable influences the model's behaviour. In regression tasks, as in the case of TS-GPR, traditional coefficients or weights do not suffice to explain how variables contribute to predictions. Consequently, a family of model-agnostic inspection methods has emerged, allowing researchers and practitioners to better interpret and diagnose a model's predictions. Three such methods available in Scikit-learn are Permutation Feature Importance, Partial Dependence Plots (PDP), and Individual Conditional Expectation (ICE) Plots.

Permutation Feature Importance is a technique that quantifies the importance of each variable by examining the deterioration in the model's performance when the values of that variable are shuffled [36]. More concretely, after training a regression model (e.g., a TS-GPR), the variable values for a predictor are permuted (shuffled) across the testing (or validation) dataset, while all other variables remain unchanged. The model's performance (e.g., R^2 or mean absolute error) is then recalculated. If shuffling a particular variable causes a large drop in performance, that variable is deemed highly important. Because this method does not rely on the model's internal parameters, it can be applied to any estimator and is thus considered model-agnostic. By comparing performance metrics before and after each permutation, one can rank the variables from most to least important. This makes it easy to identify key drivers behind the model's predictions, which can help in refining the variable set or diagnosing issues like data leakage.

Partial Dependence Plots (PDP) illustrate how the average predicted response of a model changes with respect to one or two variables of interest, while all other variables are held constant at their average values or specific representative values [37]. Specifically, the partial dependence function marginalizes over the distribution of all remaining variables, thereby revealing the average effect of a variable on the predicted outcome. For a regression model, the vertical axis of a PDP reflects the average predicted value of the target variable, and the horizontal axis indicates the range of the variable. These visual insights can guide hypotheses about the underlying relationships in the data and the model's learned behaviour.

While PDPs display the average effect of a variable, Individual Conditional Expectation (ICE) Plots provide a more granular view by showing how each individual data point's prediction changes when a single variable is varied over its range [38]. Rather than collapsing all instances into an overall curve, ICE plots create a curve per data instance, thus highlighting potential heterogeneous effects and interactions that may be obscured in an averaged plot. By visually inspecting ICE curves, one can detect whether certain subgroups of data behave differently from the average trend. This can be especially valuable in more complex regression settings where interactions between variables can significantly influence individual predictions.

In this study, we used the Scikit-learn implementations of these inspection techniques (Permutation Feature Importance, PDPs, and ICE plots) exclusively for the TS-GPR model trained in step 4 (hydrological variables R, Q_i, and Q_g with shifting plus well-correlated hydrochemical variables). These techniques were implemented using the *sklearn.inspection.permutation_importance* function for feature importance estimation, *sklearn.inspection.partial_dependence* for Partial Dependence Plots (PDPs), and *sklearn.inspection.PartialDependenceDisplay* for Individual Conditional Expectation (ICE) plots. The results were visualized using Matplotlib and Seaborn.

2.2.5. Model Validation

Lastly, the TS-GPR model was evaluated by performing inference on new data obtained after the year 2016 up to the present. The validation data consisted of 56 simulations of hydrological variables R, Q_i, and Q_g, their corresponding shifted variables with a time lag of 10 days, and 56 observations of hydrochemical variables K, Ca, Mg, Cl, and NO_3^-, all of them at irregular dates between 23 November 2023 and 30 August 2024. To measure the performance of the models, and due to the nature of the data that cover several length scales, the metrics of accuracy used in this study were the coefficient of determination (R^2), the normalized root-mean-square error (NRMSE), and the normalized mean absolute error (NMAE), which are given by:

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{N}(y_i - \overline{y})^2} \tag{4}$$

$$NRMSE = \frac{RMSE}{y_{max} - y_{min}} \tag{5}$$

$$NMAE = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_{max} - y_{min}} \tag{6}$$

where N is the number of samples, y_i is the true value, $\hat{y}_i$ is the predicted value, $\overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$ is the mean of the true values, y_{max} and y_{min} are the maximum and minimum values of the true data, and $RMSE$ is the root-mean-square error given by:

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{7}$$

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{8}$$

The TS-GPR modelling methodology employed in this study is depicted in Figure 5.

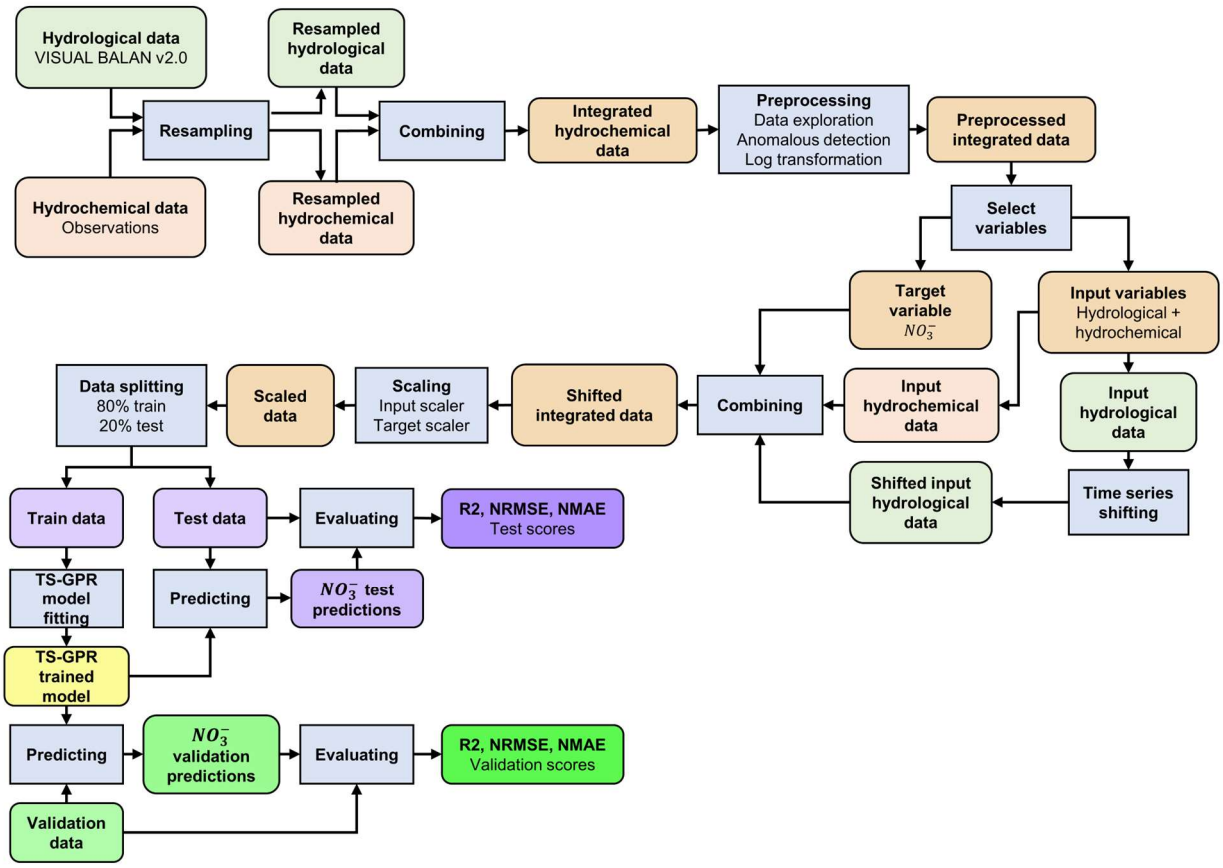


Figure 5. Flowchart of the methodology for the TS-GPR modelling in this study.

3. Results

3.1. CA Results

The optimal number of clusters was determined based on the elbow method plot (Figure 6). The elbow of the curve where the slope of the curve visibly bends from the high to low slope allows the identification of the optimal number of clusters, which in this case is equal to 3.

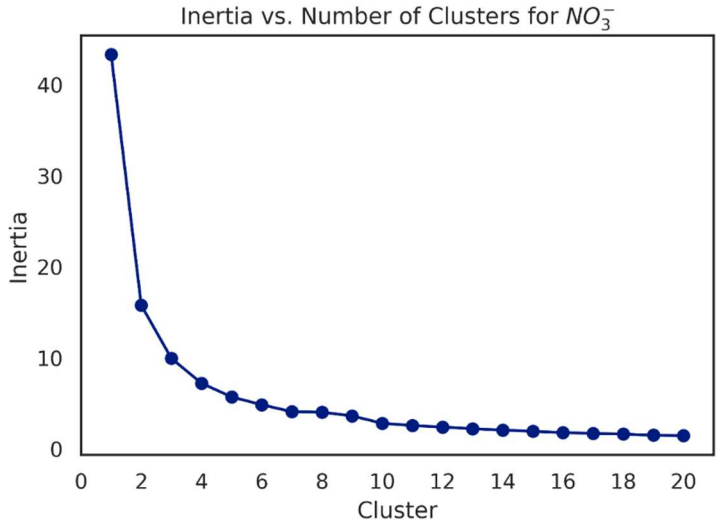


Figure 6. Plot of inertia versus number of clusters for 20 K-means models fitted on Q_1, Q_2, and nitrate data. The optimal number of clusters (three) corresponds to the elbow of the curve where the slope of the curve visibly bends from the high to low slope.

Cluster #0 in Figure 7a corresponds to the combination of high NO_3^- concentrations and low values of groundwater flow fraction, Q_2. The second cluster, cluster #1, includes the highest values of Q_2 ($Q_2 > 0.6$). The third cluster, cluster #2 comprises intermediate values of Q_2 in between the two other clusters. It is noted that the three clusters appear to be linearly separable. The results of CA were also visualized with a time series plot of the clustered NO_3^- concentration data (Figure 7b). Over time, nitrate concentrations for all clusters seem to fluctuate, but there is a notable decrease in concentration in the later years (around 2015–2017). The timing of lower NO_3^- concentrations for clusters #1 and #2 seems to coincide with a clear trend of lower values after 2012. This could be consistent with the fact that eucalyptus felling was carried out in the study area at that time, such that in the first years after the felling, there would be fewer contributions of organic matter and less mineralization. Cluster #0 maintains relatively higher levels, and its data points are dispersed over the entire timeline, suggesting greater variability in the nitrate concentrations compared to clusters #1 and 2#. This indicates that cluster #0 likely corresponds to a group of data points with more persistent higher nitrate concentrations, while clusters #1 and #2 correspond to periods or conditions with generally lower concentrations, potentially related to different environmental conditions.

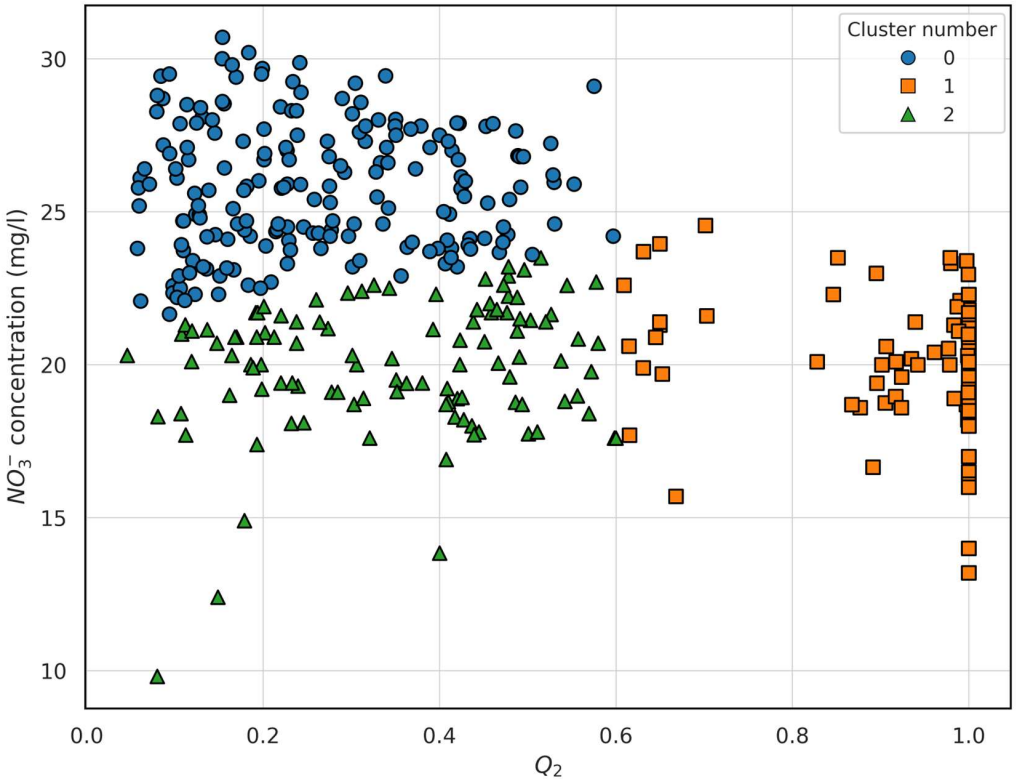
The total streamflow leaving the basin has three major components: overland flow (surface runoff), shallow subsurface flow (interflow), and groundwater flow. The results of the model show that (1) surface runoff contributes the least to total flow and its contribution is concentrated in pulses occurring after rainfall episodes; (2) interflow is the most important component, moving through the shallowest layer of weathered rock and with interflow recession lasting for about 10–15 days after rainfall; and (3) groundwater flow is the second component of the total flow, becoming the base flow during the dry season.

Nitrogen mineralization (i.e., the conversion of organic nitrogen into nitrate) is mediated by microorganisms and consists of two steps, namely ammonification and nitrification. The amount of organic matter present in the soil and the soil mineral characteristics are two factors that influence the rate of mineralization. Soil physical factors that notably affect the process are texture, pH, temperature, and water content and aeration. Nitrate export at the catchment scale is mainly dominated by transport-limited, i.e., the export of nitrate is regulated by the hydrological transport capacity. Surface runoff and interflow subsurface runoff represent the main vectors of nitrate transfer from sources to the stream.

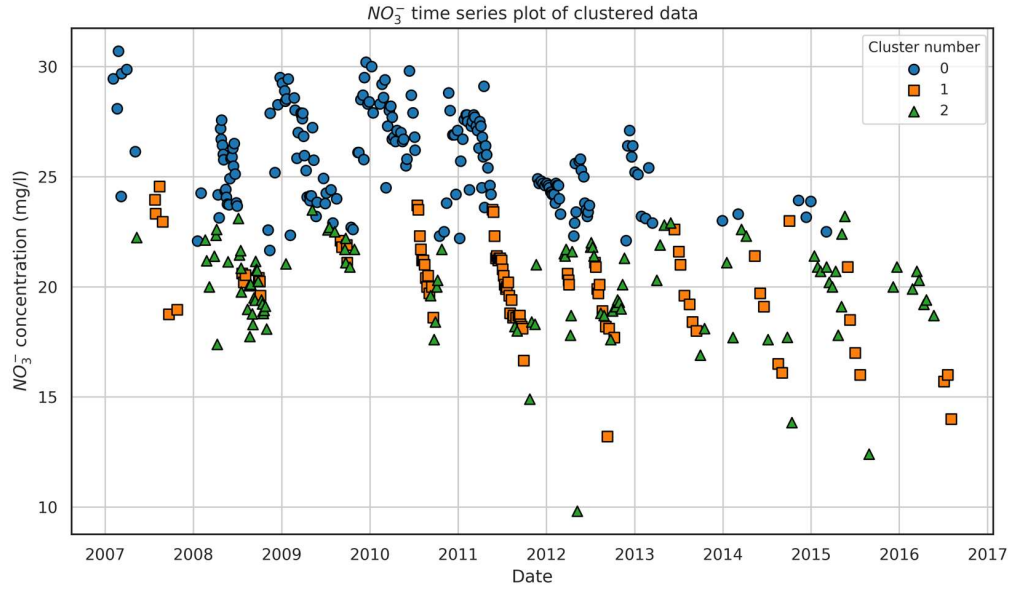
Clearly, cluster #1 corresponds to the summer period during which groundwater is the dominant component of the basin outflow (mean flow ratio = 0.026). Mineralization, and therefore nitrate concentrations, in this period are smallest because the upper soil layers at the unsaturated zone are dry, limiting bacterial growth, even if the temperatures are most adequate for microorganisms involved in nitrification. Clusters #0 and #2 correspond to periods of predominance of interflow. However, two different subperiods of interflow can be distinguished. The first one with largest flows (mean flow ratio = 0.793) is associated with the largest nitrate concentrations and corresponds to cluster #0, while the second interflow period has smaller flows (mean flow fraction = 0.548), smaller nitrate concentrations, and corresponds to cluster #1 (Tables 2–4). Therefore, the higher the interflow, the higher the nitrate concentration at the catchment outlet.

Although the environmental conditions affecting the mineralization of nitrogen (i.e., the formation of nitrates) may be very diverse, the three clusters could be roughly associated with the soil water content. Thus, cluster #0 would correspond to soil at or near field capacity in the wet season, particularly in springtime (high nitrogen mineralization); these conditions are optimal for mineralization. Cluster #2 would be associated with close-to-optimal soil water content in the wet season (nitrogen mineralization quite limited by factors such as soil temperature, or excessive soil water content, etc.). Hence, the higher

nitrate levels of cluster #0 compared to cluster #2 could be due to suboptimal soil water content or temperature of the latter. Cluster #1 would correspond to the dry season (limited nitrogen mineralization due to water deficit). Therefore, cluster analysis suggests that at the Abelar catchment nitrate flux generally increases with increasing water content in the unsaturated zone, which can be associated to increasing connectivity between upland nitrate production areas and the stream at the catchment outlet.



(**a**)



(**b**)

Figure 7. (**a**) Pair plot of nitrate concentration (mg/L) versus groundwater flow fraction Q_2. (**b**) Time series plot of nitrate concentration (mg/L) clustered data. The plot shows the three clusters identified by K-means.

Table 2. Descriptive statistics of the CA clusters across nitrate data.

	Cluster Number		
Statistic	0	1	2
count	179	76	133
mean	25.156	19.802	21.333
std	2.860	2.037	3.296
min	17.700	13.200	9.809
25% quantile	23.114	18.674	19.099
50% quantile	25.399	20.049	21.400
75% quantile	27.500	21.124	23.779
max	30.699	23.499	29.099

Table 3. Descriptive statistics of the CA clusters across Q_1 (ratio of interflow to total flow) data.

	Cluster Number		
Statistic	0	1	2
count	179	76	133
mean	0.793	0.026	0.548
std	0.093	0.045	0.116
min	0.538	0	0.297
25% quantile	0.730	0	0.479
50% quantile	0.804	0	0.545
75% quantile	0.871	0.026	0.599
max	0.942	0.171	0.919

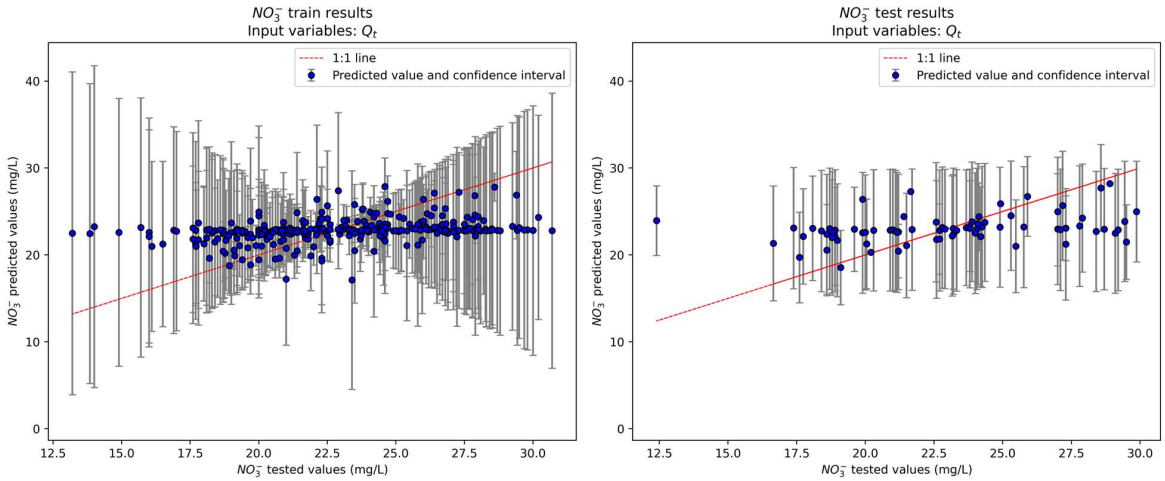
Table 4. Descriptive statistics of the CA clusters across Q_2 (ratio of groundwater flow to total flow) data.

	Cluster Number		
Statistic	0	1	2
count	179	76	133
mean	0.205	0.973	0.447
std	0.093	0.045	0.120
min	0.057	0.828	0.046
25% quantile	0.127	0.973	0.398
50% quantile	0.195	1	0.452
75% quantile	0.269	1	0.515
max	0.460	1	0.703

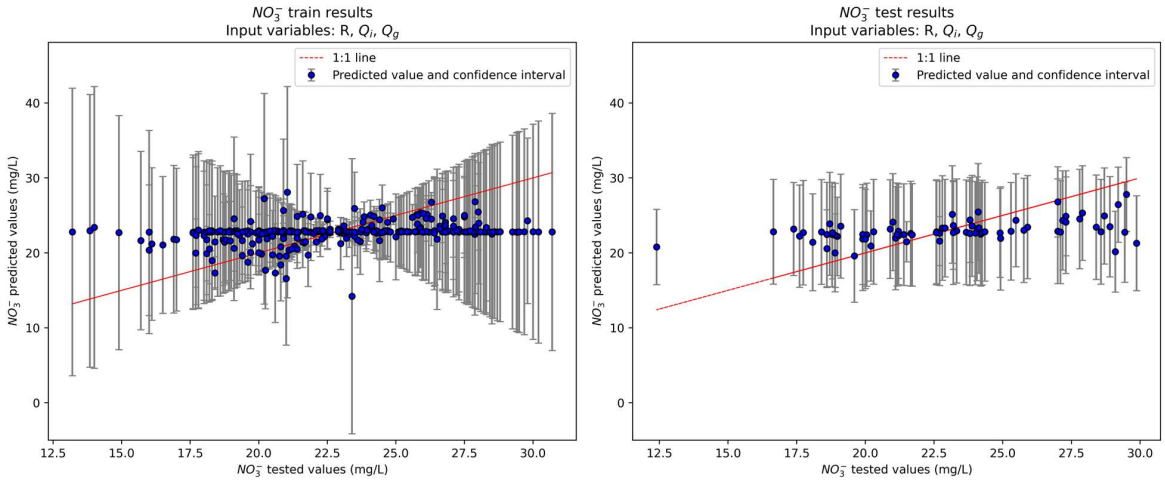
Additionally, descriptive statistics were calculated for each of the three clusters across each variable's data (Tables 2–4). The distribution of the clusters across the 388 data samples is 46.1%, 19.5%, and 34.2% for clusters #0, #1, and #2, respectively. This percentage distribution among groups is the automatic outcome of the K-means model. It is observed that cluster #0 is the most abundantly represented, containing nearly half of the dataset. Again, this suggests that, for most of the year, the soil is at field capacity. Extreme dry or wet periods are occasional.

3.2. TS-GPR Results

The baseline model uses the standard method of predicting nitrate concentration based solely on the total streamflow Q_t without shifting, which is equivalent to a time lag of 1 day. This model leads to very low R^2 scores (0.07 for training and 0.13 for testing), high NRMSE and NMAE scores, and high prediction uncertainties (Figure 8a). Testing scores improve when the best correlated hydrological variables, R, Q_i, and Q_g are considered. However, the training scores and the prediction capability remain extremely poor (Figure 8b). Results significantly improve with marginally acceptable training and testing scores and lower prediction uncertainties when hydrological variables are shifted using a time lag of 10 days (Figure 9a). Finally, exploratory analysis of all data with correlation maps was used to search for well-correlated chemical variables to include them as additional input variables. The chemical concentrations of K, Ca, Mg, and Cl were selected to predict nitrate concentration. It should be noted that no shifting was applied to these chemical variables, while shifting was applied exclusively to the hydrological variables. The results significantly improved by incorporating the chemical variables into the model, leading to R^2 = 0.82 and 0.80 for training and testing scores, respectively (Figure 9b).

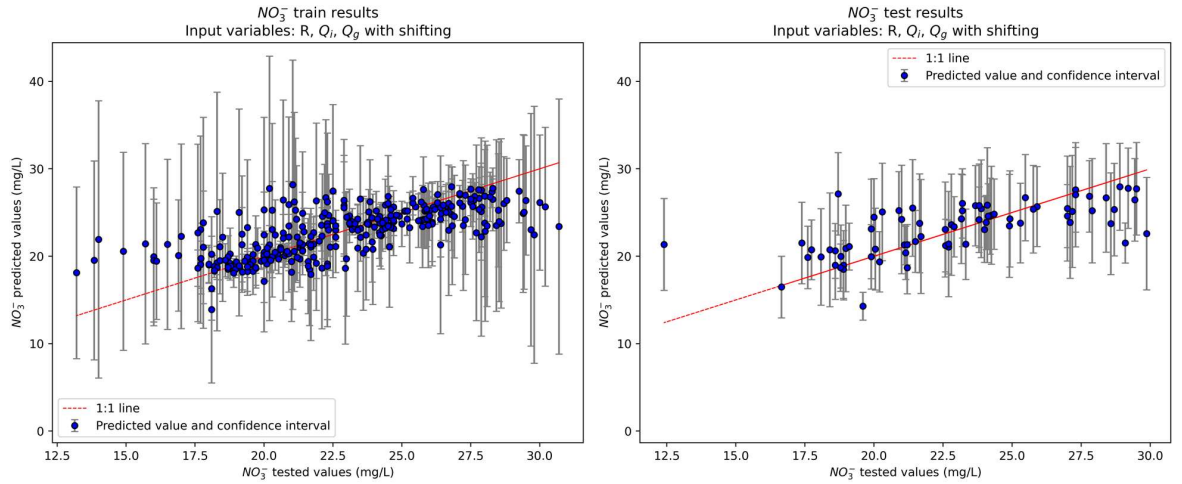


(**a**)

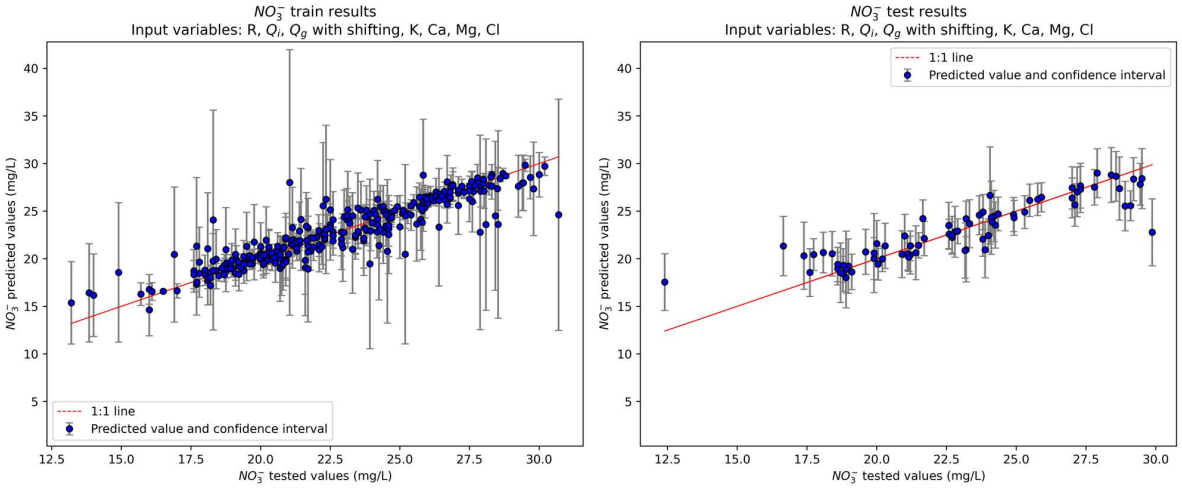


(**b**)

Figure 8. Scatter plots of predicted versus tested nitrate concentrations on the training (left plots) and testing sets (right plots) for (**a**) the baseline model, and (**b**) the model with the best correlated hydrological variables. The training results are based on k-fold cross-validation of the training set with 10 folds and 10 repeats.



(**a**)



(**b**)

Figure 9. Scatter plots of predicted versus tested nitrate concentrations on the training (left plots) and testing sets (right plots) for (**a**) the model with the best correlated hydrological variables and shifting (10 days), and (**b**) the model with the best correlated hydrological variables, shifting (10 days) and chemical variables. The training results are based on k-fold cross-validation of the training set with 10 folds and 10 repeats.

The results of this model were also visualized with a time series plot of the predicted versus tested nitrate concentrations on the testing set (Figure 10). The TS-GPR model fails to reproduce the measured NO_3^- concentrations smaller than 17.5 mg/L. The model tends to underpredict the high NO_3^- concentrations and overpredict the smallest concentrations.

The model was inspected using permutation feature importance, PDP plots, and ICE plots (Figures 11–13). The boxplot of permutation feature importances shows the relevance of the time series approach, with groundwater flow five days before the current date (Q_{g5}) being the most important variable. The concentrations of Cl, Ca, and Mg are also relevant because they contribute to greatly decrease the R^2 score when shuffled. The concentration of K also contributes to the model's performance, but with smaller relevance. The rest of the input variables might be candidates for removal or further investigation to simplify the model without losing predictive power. The boxplot also shows some outlier points outside the whiskers for variables Q_{g5}, Ca, and Mg. This suggests that in one or a few instances, the permutation of these variables led to an especially small decrease in the R^2

score compared to the average. The PDP plots in Figure 11 show an interaction between Q_{g5} and Cl. Both input variables correlate positively with nitrate concentration. The same can be said for the concentrations of Ca and Mg.

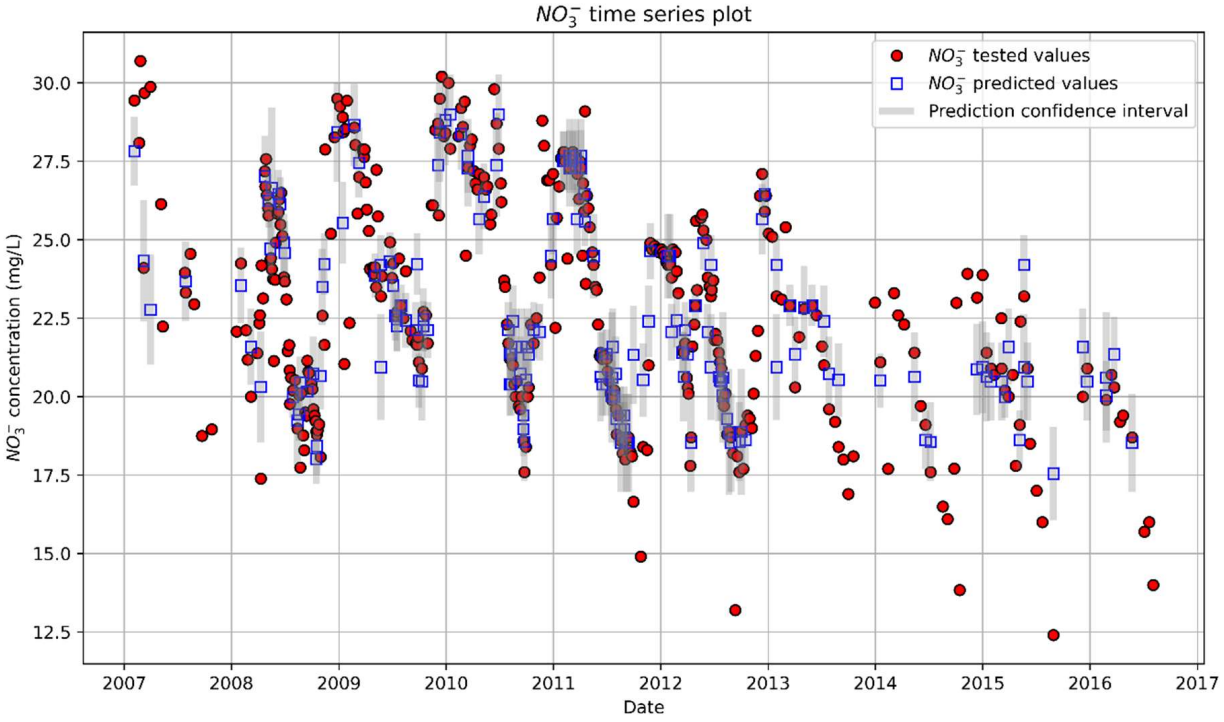


Figure 10. Time series plot of predicted versus tested nitrate concentrations. The plot illustrates the close alignment between observed and predicted nitrate concentrations over time, along with moderately narrow confidence intervals.

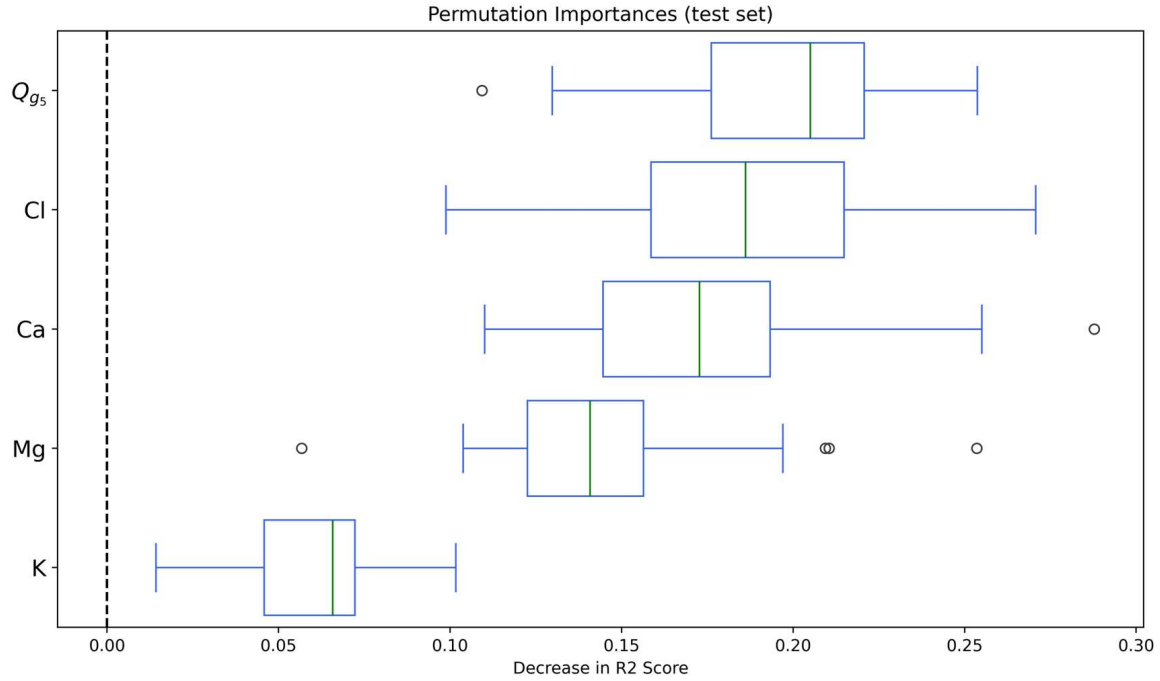
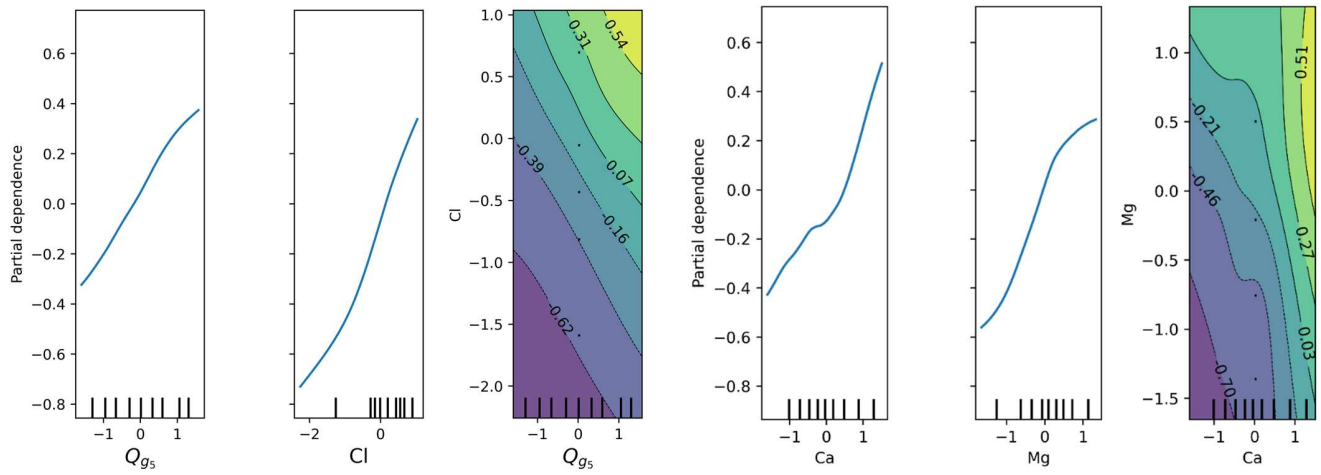
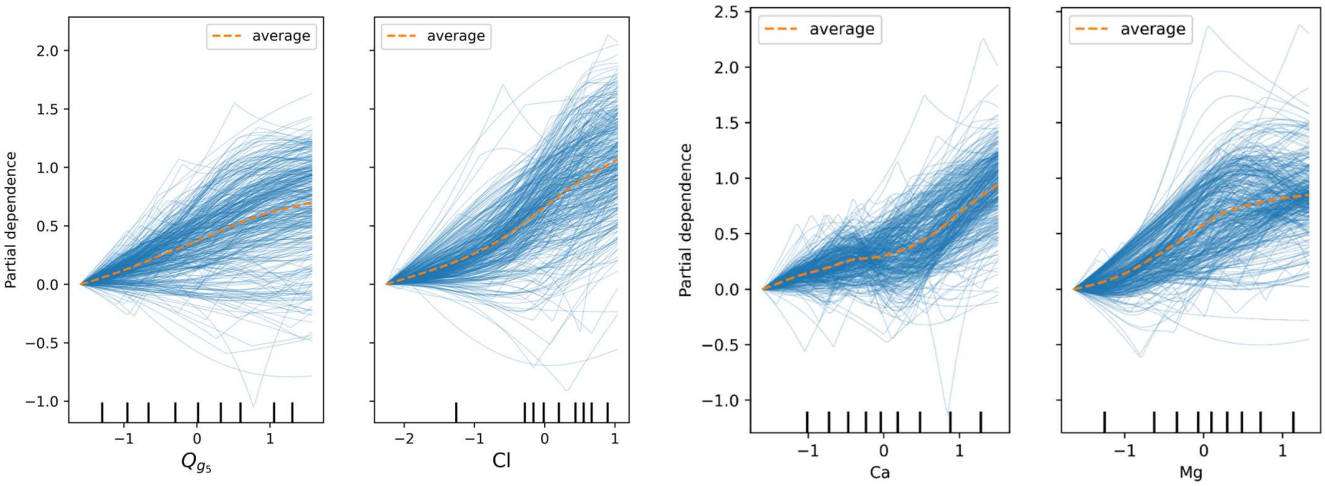


Figure 11. Boxplot of permutation features importance on the testing set. The plot shows the relevance of the time series approach with groundwater flow five days before the current date (Q_{g5}) as the most important variable. The concentrations of Cl, Ca, and Mg are also relevant as they cause great decreases in the R^2 score when shuffled.



(**a**) (**b**)

Figure 12. PDP plots of variables Q_{g5} and Cl (**a**), and Ca and Mg (**b**). The rightmost plots show an interaction effect between each pair of variables. All variables are unitless after scaling.



(**a**) (**b**)

Figure 13. ICE plots of variables Q_{g5} and Cl (**a**), and Ca and Mg (**b**). Both pairs of variables have a positive effect on the model's predictions. The dashed orange lines indicate that on average, increases in each pair of variables lead to higher nitrate concentrations.

ICE graphs in Figure 13 also attest that both variables (Q_{g5} and Cl) have a positive effect for the most part on the model's predictions. The dashed orange lines indicate that on average, NO_3^- concentrations increase when Q_{g5} and Cl increase.

The model was validated by performing inference on data collected from 2016 to 2024. The validation score of $R^2 = 0.85$ is consistent with training and testing scores (Figure 14). Other metrics such as NRMSE and NMAE are also consistent with the results obtained in training and cross-validation. Validation results were also visualized with a time series plot of the predicted versus tested nitrate concentrations on the validation set (Figure 15). From winter to early spring, the predictions align relatively well with the tested values. The model captures the overall trend, although there are minor mismatches where predicted nitrate concentrations either slightly overshoot or undershoot the true measured values. There is a noticeable increase in discrepancies from late spring to early summer. Predicted concentrations begin to underestimate the tested nitrate concentrations as concentrations decrease. Predictions were systematically smaller than tested concentrations in the last

part of the summer. It is noted that most of the observations registered at spring could correspond to cluster #0 of the CA (some of them potentially belonging to cluster #2), while the observations from the summer period would be associated with cluster #1.

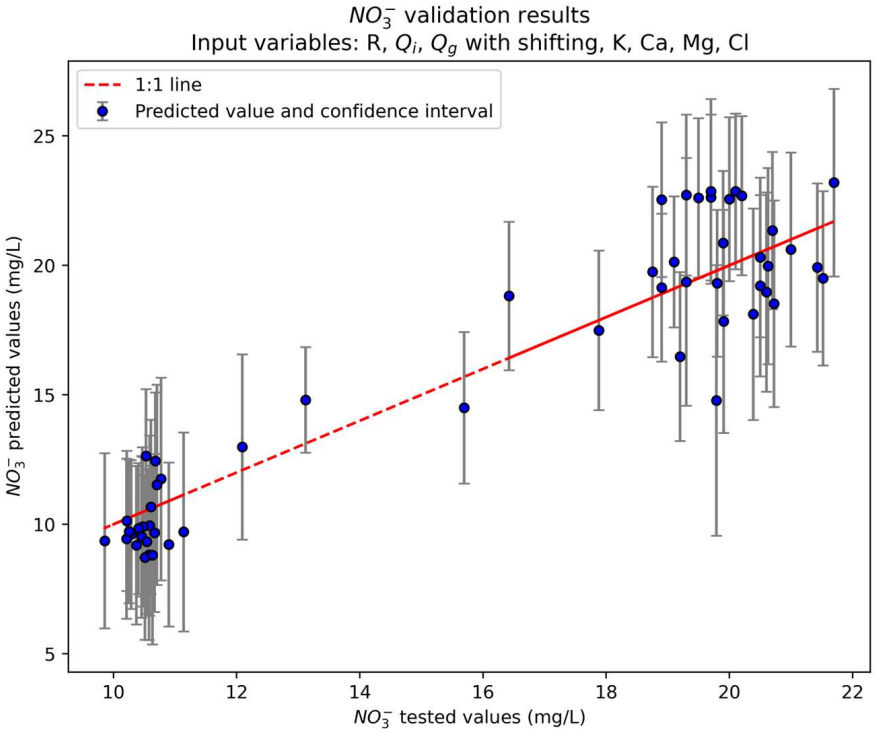


Figure 14. Scatter plot of predicted versus tested nitrate concentrations for the TS-GPR model of step 4 (with shifting of the best correlated hydrological variables and chemical variables) on the validation set.

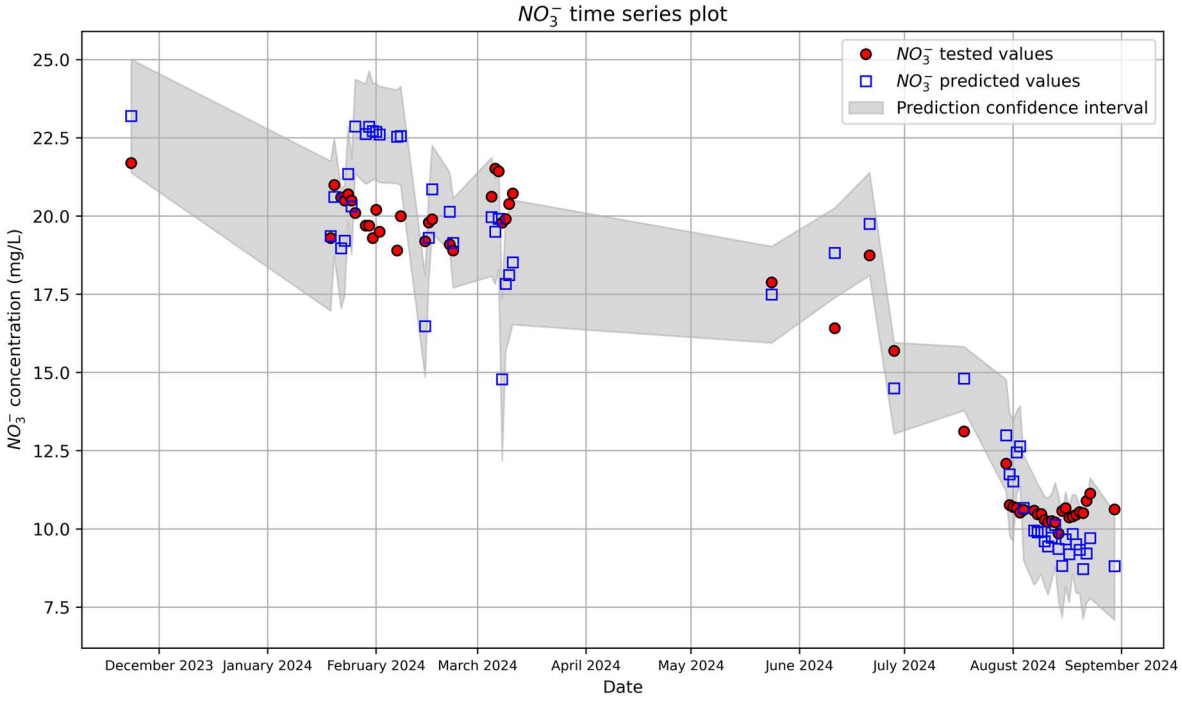


Figure 15. Time series plot of predicted versus tested nitrate concentrations on the validation set. The plot illustrates the close alignment between observed and predicted nitrate concentrations over time, along with moderately narrow confidence intervals.

The plot demonstrates that while the model performs well during periods of relative stability, it underestimates NO_3^- concentrations during periods of rapid seasonal changes, particularly in summer. This highlights the need for incorporating additional factors or revising the modelling approach to improve performance during such periods. The training, testing, and validation metrics of accuracy of the TS-GPR models at each step of the modelling process are listed in Table 5.

Table 5. Training, testing, and validation metrics of accuracy of TS-GPR models across the different steps of the modelling process. Metrics include R^2, NRMSE, and NMAE. Training metrics correspond to repeated k-fold cross-validation on the training set (10 folds, 10 repeats).

			Metrics of Accuracy		
Step	**Input Variables**	**Case**	**R^2**	**NRMSE**	**NMAE**
1	Q_t	Training	0.07	0.19	0.15
		Testing	0.13	0.20	0.15
2	R, Q_i, Q_g	Training	−0.01	0.19	0.15
		Testing	0.22	0.19	0.15
3	R, Q_i, Q_g with shifting	Training	0.44	0.14	0.10
		Testing	0.45	0.16	0.11
4	R, Q_i, Q_g with shifting and K, Ca, Mg, Cl	Training	0.82	0.08	0.05
		Testing	0.80	0.10	0.06
Validation	R, Q_i, Q_g with shifting and K, Ca, Mg, Cl	Validation	0.85	0.15	0.12

4. Discussion

Unlike previous studies (e.g., [12,13]), which focused on the use of K-means clustering to provide information on processes affecting aquifers, our research applied this method to study the relationship between a selected hydrochemical target variable (nitrate concentration in surface water) and simulated hydrological input variables (Q_1 and Q_2). K-means clustering has been extensively applied in groundwater studies to classify hydrogeochemical processes. However, its application in surface water quality modelling remains relatively limited. Recent research by Fang et al. [14] demonstrated the effectiveness of unsupervised clustering techniques in reservoir water quality assessment, highlighting the potential of classification-based methods for structuring complex hydrochemical data. Our results demonstrated that clustering successfully identified distinct hydrological subsystems, providing a structured way to classify the time variability of nitrate concentrations. These findings align with previous applications of clustering in hydrogeology but extend its use to surface water quality modelling, where classification-based methods have been used much less.

In addition, while regression models have been used very scarcely in the field of groundwater quality, and mainly through linear regression and multiple linear regression methods as the benchmark for the lowest acceptable accuracy [16], our work further explored the underrepresented regression model of GPR and extended its use to predict surface water nitrate concentration. Our results indicate that GPR provides greater flexibility in capturing nonlinear relationships between hydrological and hydrochemical variables. Compared to linear models, TS-GPR demonstrated similar predictive accuracy, particularly when time-lagged hydrological variables were incorporated. Recent studies such as that of Wang et al. [19] have shown that GPR can effectively model nonstationary hydrological datasets, demonstrating its ability to capture complex relationships in hydrological systems.

These findings suggest that TS-GPR could serve as a valuable alternative to traditional regression approaches in hydrological modelling.

The primary advantage of CA is two-fold. First, it enables the integration of simulated hydrological data with observed surface water hydrochemical data. Second, CA classifies the integrated hydrological data for selected variables into hydrological sub-systems that exhibit linear separability. Each sub-system corresponds to one of the clusters identified by the K-means algorithm, where the optimal number of clusters is determined using the elbow method. CA data provide meaningful insights, such as quantitative relationships between the variables of interest, which can be useful for identifying hydrological patterns influencing nitrate concentration. Furthermore, the cluster classification offered by CA can serve as an additional preprocessing step for hydrological data, facilitating further modelling and improving interpretability.

The main advantage of TS-GPR is the low number of required variables, which greatly reduces the chances of model overfitting. Another benefit of this approach is that it minimizes the necessary hydrochemical input variables, which are often costly to obtain in real-world scenarios. These variables could potentially be reduced even further through feature selection based on insights gained from model inspection. Similarly, TS-GPR relies on hydrological input variables that can be simulated efficiently and at low cost, making it more accessible for large-scale applications.

However, the approach also presents several challenges and limitations. The primary drawback of TS-GPR is its high computational cost, particularly in terms of CPU time and memory requirements when training models with large datasets, especially with $N > 10{,}000$ points. Future studies could explore the use of GPU acceleration or distributed computing to enhance training efficiency. Another limitation is the need for careful selection of time lags, as extending the time lag proved beneficial for model performance. Recent research by Susiluoto et al. [20] introduced a computationally efficient multi-scale Gaussian process framework capable of handling massive datasets, thereby addressing scalability issues inherent in traditional GPR models. Their findings suggest that optimizing computations and partitioning problems into parallel local subproblems can significantly enhance TS-GPR's performance, particularly in high-dimensional hydrological datasets. Future research could also investigate automated lag selection techniques, such as cross-validation–based lag optimization or attention-based methods, to further refine predictive accuracy.

Overall, this study highlights the potential of combining unsupervised learning (CA) with nonlinear regression (TS-GPR) to advance hydrological and hydrochemical modelling. Unlike traditional approaches that rely heavily on predefined assumptions about data distribution or linear relationships, our framework offers an adaptive, data-driven methodology capable of capturing complex hydrological dependencies. By integrating time-lagged hydrological variables within TS-GPR, the model leverages past hydrological conditions to improve surface water quality predictions, reducing reliance on costly hydrochemical measurements. Additionally, the use of CA as a preprocessing step for hydrological classification provides a structured way to analyse hydrochemical variations, improving both model interpretability and efficiency. These methodological advancements contribute to a scalable, cost-effective, and data-driven approach to water quality assessment. Beyond its application to nitrate modelling, this framework has the potential to be generalized to other hydrochemical parameters (e.g., manganese, sulphate, heavy metals) and extended to hybrid machine learning models for further improvements. By addressing both classification and regression challenges in hydrology, our study provides new insights into predictive water quality modelling, ultimately facilitating better environmental decision-making.

5. Conclusions

The ML analysis of hydrological and hydrochemical time series data of the Abelar basin has been presented. CA and TS-GPR methods were applied to group and predict nitrate concentration of surface water at the outlet of the basin. CA identified three approximately linearly separable clusters for the integrated hydrological system. The data samples are distributed unevenly across the three clusters, with a cluster containing the largest portion, another with a moderate portion, and the third the smallest. Cluster #0 is the most abundantly represented, containing close to half of the dataset. Cluster #0 includes data points with consistently higher nitrate concentrations, while clusters #1 and #2 are associated with periods or conditions of smaller nitrate concentrations. These patterns reflect varying environmental factors that influence nitrate concentration.

TS-GPR analysis modelled nitrate concentration based on shifted hydrological variables and selected hydrochemical variables with $R^2 = 0.82$ and 0.80 for training and testing scores and low prediction uncertainty. Model inspection confirmed the relevance of the time series approach as the most important input variable identified was Q_g from five days prior to the current date (Q_{g5}). The concentrations of Cl, Ca, and Mg were also important. The rest of the input variables are candidates for removal. Interaction effects between Q_{g5} and Cl, and between Ca and Mg, were found. The model was validated on data collected from 2016 to 2024 with results consistent with those from training and testing ($R^2 = 0.85$ for validation) demonstrating strong predictive performance and generalization. It can be concluded that the TS-GPR model achieved high performance while requiring only a limited number of hydrochemical input variables.

Future work should be devoted to using more sophisticated clustering algorithms in CA such as mean-shift, hierarchical clustering and DBSCAN, which can automatically identify the number of clusters, handle complex cluster shapes, offer greater flexibility in distance metrics and parameters, and handle outliers. Efforts should also focus on addressing the TS-GPR model's difficulty in reproducing the smallest NO_3^- concentration values. This could involve exploring the use of more flexible kernels, hyperparameter optimization, and regularization, as well as incorporating techniques such as imputation to stabilize the data and decomposition to isolate and explicitly model seasonal effects. Attention should be given to improving the performance of the relative metrics by reducing the number of input variables to only the most important ones and reevaluating the model. Exploring even greater time lags could be beneficial. The confidence intervals of the output estimates could also be leveraged to assess prediction reliability or guide further model adjustments. Additionally, subsequent research could focus on the coupling of CA and TS-GPR approaches to create a combined method, commonly known as an ensemble-based method in ML. In this way, the cluster classification offered by CA could be leveraged as an additional preprocessing step for the hydrological data, multiple TS-GPR models could be trained for each CA-identified cluster, and the predictions of each TS-GPR model would be combined to improve the performance compared to a single TS-GPR model trained on the entire dataset.

Author Contributions: Conceptualization, J.S.-C., A.M. and A.P.-G.; methodology, J.S.-P., A.M. and B.P.; software, J.S.-P. and B.P.; validation, J.S.-P. and A.P.-G.; formal analysis, J.S.-C. and A.M.; investigation, J.S.-P., B.P. and A.P.-G.; resources, J.S.-C. and B.P.; data curation, A.M., B.P. and A.P.-G.; writing—original draft preparation, J.S.-P. and J.S.-C.; writing—review and editing, J.S.-C., A.M. and A.P.-G.; visualization, J.S.-P., A.M. and B.P.; project administration, J.S.-C. and A.P.-G.; funding acquisition, J.S.-C. and A.P.-G. All authors have read and agreed to the published version of the manuscript.

Funding: The research leading to these results was funded by Project TED2021-130315B-I00 of the Strategic Plan for the Ecological and Digital Transition (EU Next Generation), the Spanish Ministry of Science and Innovation (PID2023-153202OB-I00), and the Galician Regional Government (Grant ED431C2021/54).

Data Availability Statement: Data and model results are available upon requests to the authors.

Acknowledgments: We would like to express our sincere gratitude to our colleagues Aitor García Tomillo, Acacia Naves García-Rendueles, Sara Martínez Picardo, Marcos Lado Liñares, and Brais Sobral from the AQUATERRA research group enrolled in CICA (UDC), for their dedication to on-site data collection and fieldwork and review of some parts of this manuscript, which were integral to the success of this study, and their enriching insights during the analysis phase. We are also thankful to Changbing Yang for his valuable support on ML methods and software. We acknowledge the comments and corrections of three anonymous reviewers that contributed to improve the manuscript.

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

References

1. Stylianoudaki, C.; Trichakis, I.; Karatzas, G.P. Modeling Groundwater Nitrate Contamination Using Artificial Neural Networks. *Water* **2022**, *14*, 1173. [CrossRef]
2. Castrillo, M.; García, Á.L. Estimation of High Frequency Nutrient Concentrations from Water Quality Surrogates Using Machine Learning Methods. *Water Res.* **2020**, *172*, 115490. [CrossRef] [PubMed]
3. Xu, J.; Xu, Z.; Kuang, J.; Lin, C.; Xiao, L.; Huang, X.; Zhang, Y. An Alternative to Laboratory Testing: Random Forest-Based Water Quality Prediction Framework for Inland and Nearshore Water Bodies. *Water* **2021**, *13*, 3262. [CrossRef]
4. Rodriguez-Galiano, V.; Mendes, M.P.; Garcia-Soldado, M.J.; Chica-Olmo, M.; Ribeiro, L. Predictive Modeling of Groundwater Nitrate Pollution Using Random Forest and Multisource Variables Related to Intrinsic and Specific Vulnerability: A Case Study in an Agricultural Setting (Southern Spain). *Sci. Total Environ.* **2014**, *476–477*, 189–206. [CrossRef]
5. Kratzert, F.; Klotz, D.; Brenner, C.; Schulz, K.; Herrnegger, M. Rainfall–Runoff Modelling Using Long Short-Term Memory (LSTM) Networks. *Hydrol. Earth Syst. Sci.* **2018**, *22*, 6005–6022. [CrossRef]
6. Zhi, W.; Feng, D.; Tsai, W.-P.; Sterle, G.; Harpold, A.; Shen, C.; Li, L. From Hydrometeorology to River Water Quality: Can a Deep Learning Model Predict Dissolved Oxygen at the Continental Scale? *Environ. Sci. Technol.* **2021**, *55*, 2357–2368. [CrossRef]
7. Barzegar, R.; Aalami, M.T.; Adamowski, J. Short-Term Water Quality Variable Prediction Using a Hybrid CNN–LSTM Deep Learning Model. *Stoch. Environ. Res. Risk Assess.* **2020**, *34*, 415–433. [CrossRef]
8. Bu, J.; Liu, W.; Pan, Z.; Ling, K. Comparative Study of Hydrochemical Classification Based on Different Hierarchical Cluster Analysis Methods. *Int. J. Environ. Res. Public Health* **2020**, *17*, 9515. [CrossRef]
9. Zhu, Y.; Yang, H.; Xiao, Y.; Hao, Q.; Li, Y.; Liu, J.; Wang, L.; Zhang, Y.; Hu, W.; Wang, J. Identification of Hydrochemical Characteristics, Spatial Evolution, and Driving Forces of River Water in Jinjiang Watershed, China. *Water* **2024**, *16*, 45. [CrossRef]
10. Liu, H.; Yang, J.; Ye, M.; James, S.C.; Tang, Z.; Dong, J.; Xing, T. Using *t*-Distributed Stochastic Neighbor Embedding (*t*-SNE) for Cluster Analysis and Spatial Zone Delineation of Groundwater Geochemistry Data. *J. Hydrol.* **2021**, *597*, 126146. [CrossRef]
11. Khandelwal, A.; Xu, S.; Li, X.; Jia, X.; Stienbach, M.; Duffy, C.; Nieber, J.; Kumar, V. Physics Guided Machine Learning Methods for Hydrology. *arXiv* **2020**, arXiv:2012.02854.
12. Aris, A.Z.; Praveena, S.M.; Abdullah, M.H.; Radojevic, M. Statistical Approaches and Hydrochemical Modelling of Groundwater System in a Small Tropical Island. *J. Hydroinform.* **2011**, *14*, 206–220. [CrossRef]
13. Fabbrocino, S.; Rainieri, C.; Paduano, P.; Ricciardi, A. Cluster Analysis for Groundwater Classification in Multi-Aquifer Systems Based on a Novel Correlation Index. *J. Geochem. Explor.* **2019**, *204*, 90–111. [CrossRef]
14. Fang, M.; Lyu, L.; Wang, N.; Zhou, X.; Hu, Y. Application of Unsupervised Clustering Model Based on Graph Embedding in Water Environment. *Sci. Rep.* **2023**, *13*, 22774. [CrossRef]
15. He, Q.; Barajas-Solano, D.; Tartakovsky, G.; Tartakovsky, A.M. Physics-Informed Neural Networks for Multiphysics Data Assimilation with Application to Subsurface Transport. *Adv. Water Resour.* **2020**, *141*, 103610. [CrossRef]
16. Haggerty, R.; Sun, J.; Yu, H.; Li, Y. Application of Machine Learning in Groundwater Quality Modeling—A Comprehensive Review. *Water Res.* **2023**, *233*, 119745. [CrossRef]
17. Bui, D.T.; Khosravi, K.; Karimi, M.; Busico, G.; Khozani, Z.S.; Nguyen, H.; Mastrocicco, M.; Tedesco, D.; Cuoco, E.; Kazakis, N. Enhancing Nitrate and Strontium Concentration Prediction in Groundwater by Using New Data Mining Algorithm. *Sci. Total Environ.* **2020**, *715*, 136836. [CrossRef]
18. Bhattarai, A.; Dhakal, S.; Gautam, Y.; Bhattarai, R. Prediction of Nitrate and Phosphorus Concentrations Using Machine Learning Algorithms in Watersheds with Different Landuse. *Water* **2021**, *13*, 3096. [CrossRef]

19. Wang, S.; Gong, J.; Gao, H.; Liu, W.; Feng, Z. Gaussian Process Regression and Cooperation Search Algorithm for Forecasting Nonstationary Runoff Time Series. *Water* **2023**, *15*, 2111. [CrossRef]
20. Susiluoto, J.; Spantini, A.; Haario, H.; Härkönen, T.; Marzouk, Y. Efficient Multi-Scale Gaussian Process Regression for Massive Remote Sensing Data with satGP v0.1.2. *Geosci. Model Dev.* **2020**, *13*, 3439–3463. [CrossRef]
21. Samper, J.; Naves, A.; Pisani, B.; Montenegro, L.; Mon, A.; Fernández, J.; Arias, R.; Piñeiro, R.; Velo, M.; Ameijenda, C. Estudio Hidrogeológico, Vulnerabilidad y Protección de Las Captaciones de Los Suministros Rurales En Abegondo (A Coruña). In Proceedings of the Conference "Las Aguas Subterráneas y la Planificación Hidrológica: Congreso Hispano-Luso Sobre las Aguas Subterráneas en el Segundo Ciclo de Planificación Hidrológica", Madrid, Spain, 28–30 December 2016; pp. 335–344.
22. Naves, A.; Samper, J.; Mon, A.; Pisani, B.; Montenegro, L.; Carvalho, J.M. Demonstrative Actions of Spring Restoration and Groundwater Protection in Rural Areas of Abegondo (Galicia, Spain). *Sustain. Water Resour. Manag.* **2019**, *5*, 175–186. [CrossRef]
23. Samper, J.; Naves, A.; Pisani, B.; Dafonte, J.; Montenegro, L.; García-Tomillo, A. Sustainability of Groundwater Resources of Weathered and Fractured Schists in the Rural Areas of Galicia (Spain). *Environ. Earth Sci.* **2022**, *81*, 141. [CrossRef]
24. Soto, B.; Brea, M.A.; Pérez, R.; Díaz-Fierros, F. Influence of 7-Year Old Eucalyptus Globulus Plantation on the Low Flow of a Small Basin. In Proceedings of the Water in Celtic Countries: Quantity, Quality and Climate Variability (Proceedings of the Fourth InterCeltic Colloquium on Hydrology and Management of Water Resources, Guimarães, Portugal, 11–14 July 2005; pp. 232–241.
25. Rodríguez-Suárez, J.A.; Soto, B.; Perez, R.; Diaz-Fierros, F. Influence of Eucalyptus Globulus Plantation Growth on Water Table Levels and Low Flows in a Small Catchment. *J. Hydrol.* **2011**, *396*, 321–326. [CrossRef]
26. Peel, M.C.; Finlayson, B.L.; McMahon, T.A. Updated World Map of the Köppen-Geiger Climate Classification. *Hydrol. Earth Syst. Sci.* **2007**, *11*, 1633–1644. [CrossRef]
27. Samper; Pisani, B.; Marques, J.E. Hydrological models of interflow in three Iberian mounntain basins. *Environ. Earth Sci.* **2025**, *73*, 2645–2656. [CrossRef]
28. Pisani, B.; Samper, J.; Espinha-Marques, J. Climate change impact on groundwater resources of a hard rock mountain region (Serra da Estrela, Central Portugal). *Sustain. Water Resour. Manag.* **2019**, *5*, 289–304. [CrossRef]
29. Espinha Marques, J.; Samper, J.; Pisani, B.; Alvares, D.; Carvalho, J.M.; Chaminé, H.I.; Marques, J.M.; Vieira, G.T.; Mora, C.; Sodré Borges, F. Evaluation of Water Resources in a High-Mountain Basin in Serra Da Estrela, Central Portugal, Using a Semi-Distributed Hydrological Model. *Environ. Earth Sci.* **2011**, *62*, 1219–1234. [CrossRef]
30. Stigter, T.Y.; Nunes, J.P.; Pisani, B.; Fakir, Y.; Hugman, R.; Li, Y.; Tomé, S.; Ribeiro, L.; Samper, J.; Oliveira, R.; et al. Comparative assessment of climate change impacts on coastal groundwater resources and dependent ecosystems in the Mediterranean. *Reg. Environ. Change* **2014**, *14* (Suppl. S1), S41–S56. [CrossRef]
31. Tukey, J.W. *Exploratory Data Analysis*; John, W., Ed.; Addison-Wesley Pub. Co.: Reading, MA, USA, 1977; ISBN 978-0-201-07616-5.
32. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
33. Macqueen, J. Some Methods for Classification and Analysis of Multivariate Observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, Davis, CA, USA, 1 January 1967; Volume 1, pp. 281–297.
34. Rasmussen, C.E.; Williams, C.K.I. *Gaussian Processes for Machine Learning*; Adaptive computation and machine learning; MIT Press: Cambridge, MA, USA, 2008; ISBN 978-0-262-18253-9.
35. Seeger, M. Gaussian Processes for Machine Learning. *Int. J. Neur. Syst.* **2004**, *14*, 69–106. [CrossRef]
36. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
37. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
38. Goldstein, A.; Kapelner, A.; Bleich, J.; Pitkin, E. Peeking Inside the Black Box: Visualizing Statistical Learning with Plots of Individual Conditional Expectation. *J. Comput. Graph. Stat.* **2015**, *24*, 44–65. [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.