ENVIRONMENTAL RESEARCH
COMMUNICATIONS

PAPER • OPEN ACCESS

Machine learning models to predict nitrate concentration in a river basin

To cite this article: Diana Yaritza Dorado-Guerra *et al* 2022 *Environ. Res. Commun.* **4** 125012

View the article online for updates and enhancements.

Environmental Research Communications

OPEN ACCESS

RECEIVED
14 October 2022

REVISED
2 December 2022

ACCEPTED FOR PUBLICATION
14 December 2022

PUBLISHED
11 January 2023

PAPER

Machine learning models to predict nitrate concentration in a river basin

Diana Yaritza Dorado-Guerra[1], Gerald Corzo-Pérez[2], Javier Paredes-Arquiola[1] and Miguel Ángel Pérez-Martín[1]

[1] Research Institute of Water and Environmental Engineering (IIAMA), Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain

[2] UNESCO-IHE Institute for Water Education, PO Box 3015, 2601DA Delft, The Netherlands

E-mail: diadogue@doctor.upv.es

Keywords: nitrate, random forest, eXtreme gradient boosting, feature selection, surface water bodies, boruta shap

Abstract

Aquifer-stream interactions affect the water quality in Mediterranean areas; therefore, the coupling of surface water and groundwater models is generally used to solve water-planning and pollution problems in river basins. However, their use is limited because model inputs and outputs are not spatially and temporally linked, and the data update and fitting are laborious tasks. Machine learning models have shown great potential in water quality simulation, as they can identify the statistical relationship between input and output data without the explicit requirement of knowing the physical processes. This allows the ecological, hydrological, and environmental variables that influence water quality to be analysed with a holistic approach. In this research, feature selection (FS) methods and algorithms of artificial intelligence—random forest (RF) and eXtreme Gradient Boosting (XGBoost) trees—are used to simulate nitrate concentration and determine the main drivers related to nitrate pollution in Mediterranean streams. The developed models included 19 inputs and sampling of nitrate concentration in 159 surface water quality-gauging stations as explanatory variables. The models were trained on 70 percent data, with 30 percent used to validate the predictions. Results showed that the combination of FS method with local knowledge about the dataset is the best option to improve the model's performance, while RF and XGBoost simulate the nitrate concentration with high performance ($r = 0.93$ and $r = 0.92$, respectively). The final ranking, based on the relative importance of the variables in the RF and XGBoost models, showed that, regarding nitrogen and phosphorus concentration, the location explained 87 percent of the nitrate variability. RF and XGBoost predicted nitrate concentration in surface water with high accuracy without using conditions or parameters of entry and enabled the observation of different relationships between drivers. Thus, it is possible to identify and delimit zones with a spatial risk of pollution and approaches to implementing solutions.

1. Introduction

Nitrate is an important component in the environment. Its availability influences food supply, water and habitat quality, while toxic effects on stream biota and human health can occur with high concentrations of nitrate (Singh *et al* 2022). Its main source in Europe is diffuse pollution (Grinsven *et al* 2015, Alcon *et al* 2022), whereby nitrogen leaches when transformed into nitrate form. The main issue, then, with nitrates is their mobility in soil, and the fact that they can persist in surface water (SW) and groundwater (GW) (Defterdarović *et al* 2021), contributing to poor water quality and eutrophication (Pang *et al* 2022). Currently, the ecological status of more than half the water bodies in the EU is assessed as poor (Poikane *et al* 2019), contrary to the requirements of the Water Framework Directive (WFD) and Nitrates Directive (Directive 91/676/EEC). Decreasing nitrate concentration is already a challenge in several areas of Europe (Grizzetti *et al* 2021, Tzilivakis *et al* 2021) with approximately 40 percent of water bodies in Spain assessed as having poor water quality (Ministerio para la transción ecológica y el reto Demográfico 2020). A similar situation occurred in the Júcar River Basin (RB) (the fourth most populated region in Spain), where 61 percent of the 124 water bodies have been assigned poor quality status, according to the local hydrological plans (Confederación Hidrográfica del Júcar 2022a).

The usual methods for the assessment of nitrate concentrations consist of numerical modelling of pollutant transport, which is an efficient tool for understanding the physical, chemical and biological processes of nitrate transport (Singh and Craswell 2021). Complete reviews of models used in pollution estimation were conducted by Bouraoui and Grizzetti 2014 and Yuan *et al* 2020, however, most models simulate rather simplified scenarios, such as a single soil type, and layered soil types in a two-dimensional vertical domain, or river–aquifer interactions are not represented. As a result of the above, the coupling of the two models PATRICAL (Pérez-Martín *et al* 2014) and RREA (Paredes-Arquiola 2021) was employed to simulate the nitrate concentration in the Júcar RB, with the combination of the models representing river-aquifer interactions, the hydrological cycle altered by humans, irrigation returns, and lateral transfer among aquifers. According to the result reported by Dorado-Guerra *et al* 2021, this coupling found 58 percent of lineal correlation between simulated and observed nitrate concentration. The heterogeneity of the study area, availability of data, and complexity of integrated SW-GW modelling means other techniques are needed to improve the accuracy and computational cost of the nitrate concentration predictions.

Recently, artificial intelligence algorithms have been applied in hydrological studies of nitrate pollution with good results. These algorithms can efficiently solve complex non-linear problems, as they learn from the dataset and therefore do not require pre-defined rules based on expert criteria (Zhu *et al* 2022). Furthermore, artificial intelligence algorithms have been found to increase predictive performance across a wide range of environmental processes (Tyralis *et al* 2019). Presently, ensemble learning such as random forest (RF) and eXtreme Gradient Boosting (XGBoost) are widely adopted in water science; however, previous investigations indicated limited application of RF and XGBoost algorithms to predict SW nitrate concentrations (Zhu *et al* 2022). Furthermore, no such studies were found looking at the Júcar RB. RF's advantages include the ability to capture non-linear dependencies and interactions of variables, computational speed, parsimonious parameterisation, and the use of variable importance metrics (Tyralis *et al* 2019). Various studies have been conducted to predict nitrate distribution patterns in GW using the RF algorithm (Rodriguez-Galiano *et al* 2014, Bao *et al* 2022, He *et al* 2022). According to Castrillo and López 2020, RF is suitable for representing the concentration of nutrients in either a rural or urban catchment. On the other hand, XGBoost can improve the model's robustness and running speed by introducing terms for regularisation, column sampling and the decision tree's ability to choose the split point (Ma *et al* 2021, Gervasi *et al* 2022). In applied water quality studies, XGBoost performed better against other algorithms such as LogiBoost, RF, AdaBoost, and support-vector machines (Garabaghi 2022, Izzuan *et al* 2022, Li *et al* 2022, Nasir *et al* 2022).

In terms of the prediction of nitrates, many factors have been reported in research studies as influential and playing a crucial role, including location, nitrogen, ammonium, phosphate, pH level, ambient and water temperature, dissolved oxygen, biological oxygen demand, suspended solids, and streamflow (Wu *et al* 2017, Bagherzadeh *et al* 2021). In order to select the most informative variables for dealing with the problem, feature selection (FS) methods need to be applied. FS removes irrelevant and noisy features while keeping those with minimum redundancy and maximum relevance to the target variable, and its application results in more cost-effective models and improves algorithm performance (Effrosynidis and Arampatzis 2021). Although there are many FS methods, most studies use correlation methods only, such as Pearson's correlation. Therefore, a comparative assessment of the effect of FS on improving the accuracy of simulating nitrate concentration in surface water is still needed.

This study aims to investigate the effect of FS methods and two artificial intelligence algorithms in terms of enhancing the prediction performance of SW nitrate concentration in water bodies of the Júcar RB. The specific objectives of this paper are fourfold: (1). Defining groups of variables according to the FS result; (2). creating AI models using algorithms such as RF and XGBoost; (3). finding the best nitrate concentration forecasting model; and (4). finding the features that most influence nitrate concentration in the Júcar RB. A total of 19 features were adopted for application of FS methods and to construct the proposed models: air temperature (T_a), precipitation, distance from the river source (DRS), streamflow, piezometric level (PL), water temperature (T_w), pH level, nitrogen (N), nitrite (NO_2), ammonium (NH_4), biochemical oxygen demand over five days (BOD_5), suspended solids (SS), dissolved oxygen (DO), total phosphorus (TP), nitrate GW, the Specific Pollution Sensitivity Index (IPS, in Spanish), the Iberian Biological Monitoring Working Party (IBMWP), the quality riparian index (QBR, in Spanish), and load of diffuse pollution (DP). The novelty of this study is its inclusion of ecological indicators and the relationship between river and aquifer with the PL and nitrate in the GW. In areas with water scarcity and high river-aquifer connectivity, such as the Júcar RB, where conjunctive use of GW and SW is typical, the contribution of the GW component to SW pollution is important.

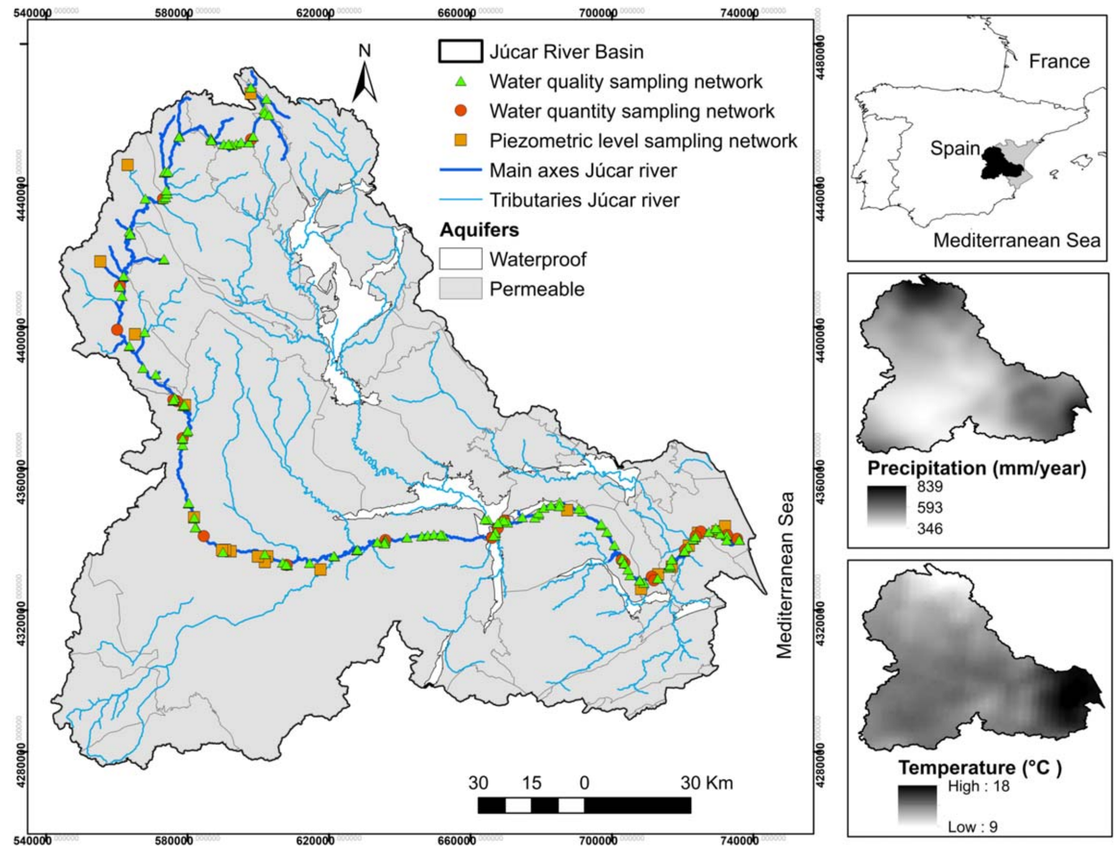

Figure 1. Location of the Júcar River Basin, water sampling network, mean annual precipitation, and mean temperature (1980–2020).

2. Material and methods

2.1. Case study

The Júcar RB is located within the Júcar River Basin District in the east of the Iberian Peninsula (Spain) on the Mediterranean side, with an area of 22,208 Km^2 (figure 1). The Júcar River has the largest catchment area and the greatest flow contribution of the Júcar RB District, with 36 surface water bodies and a length of 509 km on the main axis, which empties into the Mediterranean Sea. In the geomorphological context, the main characteristics of the basin can be grouped into two main zones: a mountainous interior, with peaks between 1,500 and 2,028 m, but which develops below 1,000 m and a second coastal zone, made up of coastal plains. This plain is an alluvial platform that provides nutrient-rich soil that supports most of the irrigated agricultural production, and is home to more than 80% of the basin's total population (Confederación Hidrográfica del Júcar 2022b).

Average temperatures range from less than 10 °C inland to 18 °C in the coastal zone (figure 1). The climate varies from humid to semi-arid, with the presence of droughts and a concentration of approximately 42 percent of the annual rainfall in autumn on the coastal strip. The average annual rainfall is 504 mm year^{-1}, with a spatial range of 797 mm year^{-1} in the headwater, 368 mm year^{-1} in the midstream and 679 mm year^{-1} at the mouth of the river at the Mediterranean Sea. The contribution to the main river network in the Júcar RB is 1245 hm^3 year^{-1} with 23.9 hm^3 year^{-1} discharging into the Mediterranean Sea. The great hydrological variability and the scarcity of resources in the basin has meant that, in order to meet the demand, especially for irrigation water, a large number of hydraulic infrastructures have been built with a total water storage capacity of 2,846 hm^3 (Confederación Hidrográfica del Júcar 2022b).

According to the dominant lithology of the GW bodies (IGME-DGA 2012), the outcrop can be classified as 25 percent detrital and 29 percent carbonate, with the rest being of mixed origin from both materials. The water bodies on the main axis of the river are classified as gaining stream (64 percent receiving discharges from the GW), losers (14 percent of the river infiltrating resources into the GW), and variable (22 percent representing one situation or another depending on the time of the year). The nitrate concentration of 25 percent of the aquifer is above the good status threshold, located in the midstream and downstream sections (Confederación Hidrográfica del Júcar 2022a).

The land use in the Júcar RB (EEA 2021) roughly breaks down into forest areas and open spaces (49 percent), agriculture (49 percent), and artificial surfaces (2 percent). Agriculture is the activity with the highest water resource requirement (85 percent of total demand), and the dry season (July and August) coincides with the most water demanding period (Ortega-Reig *et al* 2017). The water demand is 1338 hm^3 $year^{-1}$, of which 55 percent is supplied by rivers, and 41 percent by aquifers. The total rainfall area is 209,773 ha, 38 percent of which corresponds to citrus crops, located in the downstream of the basin, the area with the highest nitrate concentration in rivers and aquifers. The second and third most important groups are winter cereals for grain and the grape crop, each covering 11 percent of the area. However, net water demand is higher for rice crops (8011 m^3 ha^{-1} $year^{-1}$), while citrus requires 3890 m^3 ha^{-1} $year^{-1}$ (Confederación Hidrográfica del Júcar 2022c). Citrus orchards and rice crops with irrigation are the main sources of diffuse pollution in the basin. The largest cities in the basin are Albacete (385,000 inhabitants) and Cuenca (198.842 inhabitants). The discharge wastewater produced by domestic and industrial uses amounts to 20 hm^3 $year^{-1}$ in the two cities. The greatest load of nitrate pollution comes from agriculture rather than from point sources (Dorado-Guerra *et al* 2021).

2.2. Observed data

The variable target nitrate concentration, water quality, water quantity and ecological parameters in SW, PL, and nitrate concentration in GW were measured by the Júcar RB District authority and the dataset is available on the Water Information System for the Júcar RB District report ('SIA Júcar' in Spanish: aps.chj.es/siajucar/, accessed on March 26 2021). The different sampling networks are shown in figure 1.

T_w, pH, N, NO_2, NH_4, BOD_5, SS, DO, and TP have recently been factors used to forecast nitrate concentration using machine learning models (Latif *et al* 2020). The variable target nitrate concentration and previous parameters have been measured at surface water quality gauging stations at 159 points since 1990.

Some studies have revealed the dependent relationship between hydrological factors and nitrate concentration in SW bodies with precipitation and streamflow playing an important role in the fluctuations across different temporal scales (Gu *et al* 2020). Precipitation and T_a were acquired from AEMET (the State Meteorological Agency in Spain), which has a high-resolution (0.05 degrees) daily gridded precipitation dataset for Peninsular Spain and the Balearic Islands (version 2) (Peral García *et al* 2021). The point nearest to the surface water body was taken as the reference for precipitation in each of the river reaches, where the streamflow has been measured at 20 points since 1970. GW and SW interactions can be significant when modelling nitrate concentration in rivers in the region where piezometric levels and nitrate concentration in the GW are high (Rafiei *et al* 2022). The PL has been measured in 19 wells since 1990.

Changes are expected in the community structure after stress levels or pollutant agents and provide an early indication of possible adverse effects within the ecosystem. The Specific Pollution Sensitivity Index (IPS) measures the relative abundance of diatom species, and, with a score range from 0 to 20, the reaches with values above 18 are classified as good quality, while values close to 0 are classified as poor quality (Cemagref 1982). The Iberian Biological Monitoring Working Party (IBMWP) index is determined by the numbers of macroinvertebrate families (Alba-Tercedor *et al* 2002). Index scores range from 0 to 235 points, and reaches with values above 100 are classified as good quality, while values close to 0 are classified as poor quality. The quality riparian index (QBR) is used to assess the quality of the riparian vegetation, providing a rapid assessment of the overall condition of the riparian zone using four aspects (total riparian vegetation cover, cover structure, and quality and degree of naturalness of the stream channel). The QBR index scores range from 0 to 100, with reaches attaining values above 95 classified as good quality, and values close to 0 as poor (Munné *et al* 2003). Ecological indicators have been measured every year since 2009.

Anthropogenic effects have been taken into account when dealing with DP, which corresponds to 99 percent of the nitrate load in the Júcar RB District. DP comes from the PATRICAL model using the methodology detailed in Dorado-Guerra *et al* 2021.

Once the time series were obtained for each SW body, the median of all parameters was calculated on a quarterly scale, with the exception of temperature and precipitation, which were entered into the model as cumulative. The analysis was performed for the period between 2009 and 2019, due to ecological indicator data being available since 2009. Table 1 shows the independent parameters, including data sources and timescale (Dorado-Guerra *et al* 2022).

2.3. Methodology

In total, 19 parameters, including climatic, hydrological, hydrogeological, ecological, water quality and anthropogenic, were used as inputs for modelling the SW nitrate concentration using RF and XGBoost models. The models were calibrated and validated with 70 and 30 percent of the dataset, respectively, which consisted of the target value and prediction factors at the location of each SW body from 2009 to 2019. Records with missing values were excluded from training and test datasets. As a result, some features with only few samples were excluded and the cross-validation (CV) method was applied, which allowed the algorithm to learn from the totality of the data, so that the data was unbiased. In order to identify the best input combination for nitrate

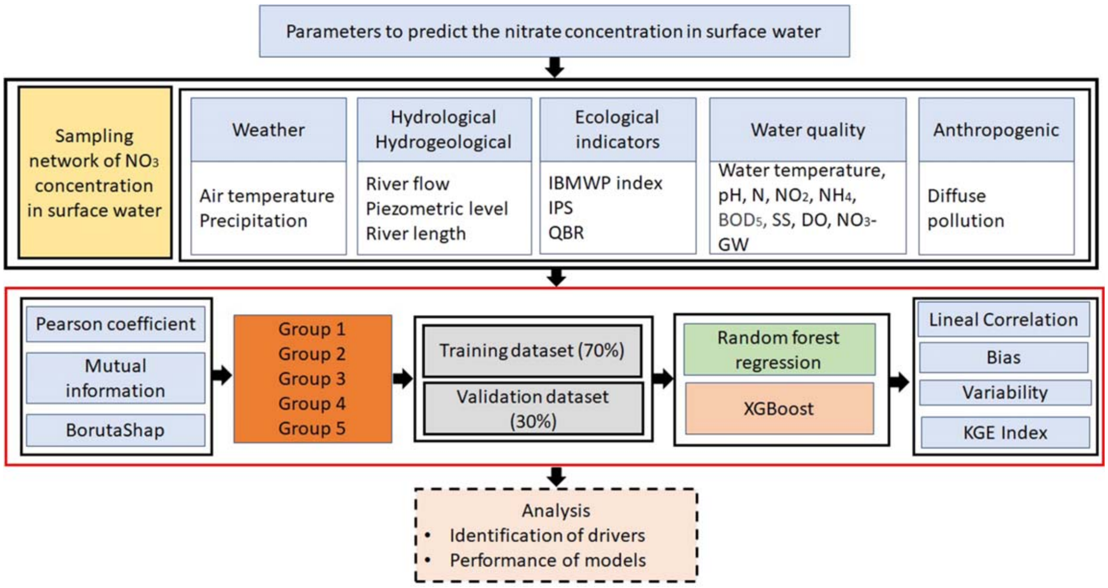

Figure 2. Steps for simulating nitrate concentration in surface water using random forest regression and XGBoost models. The abbreviations used are defined in table 1.

Table 1. Summary of the parameters and data sources.

Parameters	Abbreviation	Monitoring Points	Source
Weather			State Meteorological Agency in Spain (https://aemet.es)
Air temperature (°C)	T_a	36	
Precipitation (mm)		36	
Hydrological—Hydrogeologial			SIA Júcar (https://aps.chj.es/siajucar/)
Distance from the river source (Km)	DRS		
Streamflow (hm^3 $year^{-1}$)		20	
Piezometric level (m.a.s.l)	PL	19	
Water Quality			SIA Júcar (https://aps.chj.es/siajucar/)
Nitrate SW ($mgNO_3$/L)		159	
Water temperature (°C)	T_w	159	
pH	pH	159	
Nitrogen (mgN/L)	N	159	
Nitrite ($mgNO_2$/L)	NO_2	159	
Ammonium ($mgNH_4$/L)	NH_4	159	
Biochemical oxygen demand over five days	BOD_5	159	
Suspended solids ($mg\,l^{-1}$)	SS	159	
Dissolved oxygen (mgO_2/L)	DO	159	
Total phosphorus (mgP/L)	TP	159	
Nitrate GW ($mgNO_3$/L)			
Ecological indicators			SIA Júcar (https://aps.chj.es/siajucar/)
Specific Pollution Sensitivity Index	IPS	36	
Iberian Biological Monitoring Working Party index	IBMWP	36	
Quality riparian index	QBR	36	
Anthropogenic			Dorado-Guerra *et al* 2021
Diffuse pollution	DP	36	

estimation, a comprehensive feature selection analysis was carried out using Pearson correlation, mutual information (MI) and the BorutaShap algorithm. The applied methodology is depicted in figure 2.

2.3.1. Feature selection

When the number of inputs is high, selecting the best inputs has an important impact on the model accuracy and computational cost (Rodriguez-Galiano *et al* 2018, Effrosynidis and Arampatzis 2021). Therefore, to recognise the best input combination for estimating nitrate concentration, a feature selection analysis was carried out using Pearson correlation, MI and the BorutaShap algorithm as a random forest-based wrapper process.

MI is a measure of the quantity of information that a random variable shares with another variable. The mathematical definition of MI is described in Cover and Thomas 2006, and Vergara and Estévez 2014. It is related linearly to the entropies of variables: a nonlinear measure that can be a useful tool to determine the dominant inputs among large numbers of parameters, thereby supporting the information obtained with Pearson's coefficient. Features are ranked from largest to smallest MI values in terms of the target.

BorutaShap is a wrapper-feature selection methodology that merges the Boruta algorithm with the SHAP (Shapley Additive Explanations) framework for feature importance and ranking, and the sampling procedure uses smaller sub-samples of the available data at each iteration of the algorithm. Boruta and BorutaShap are based on a RF algorithm, which is faster than other algorithms, can usually be run without parameters tuning, can capture non-linear dependencies between predictor and dependent variables and provides a numerical score of feature importance (Kursa and Rudnicki 2010). The BorutaShap algorithm uses the following process (Keany 2021): (1). create shadow features (new copies of all the features in the dataset), and add the shadow features back to the dataset; (2). estimate the feature importance metrics of original and shadow features; (3). generate a threshold using the maximum importance score of the shadow features, and assign a hit to any features that are above the threshold; (4). carry out a two-sided t-test of equality for each unassigned feature; (5). classify the features into three groups— features with an importance significantly above the threshold ('important'), those that outperform at a less than the threshold ('tentative'), and features with an importance significantly below the threshold ('unimportant'), which are removed from the process; and 6. delete all shadow features and repeat the procedure until an importance has been assigned to each feature. The Boruta-SHAP library for Python was then applied to the feature selection (Keany 2020).

2.3.2. Machine learning models

Supervised learning algorithms, such as RF, are increasingly being used in SW pollution modelling (e.g., Thornhill *et al* 2017, Jamei *et al* 2022). RF is an assemblage of a large number of classification or regression trees, which uses a sample of the data to build a model. For regression targets, RF generates several decision trees and aggregates the predictions using bootstrapping, thereby averaging the predictions to construct a model using only a proportion of the predictors (Breiman 2001). The correlation between decision trees decreases, thereby improving the predictive power and reducing the computational complexity of the algorithm (Tyralis *et al* 2019).

XGBoost is an enhancement of the gradient-boosting decision tree algorithm (Chen and Guestrin 2016) with the main objective to improve the accuracy and speed of the model. Each update in the algorithm is based on the prediction results of the previous one; by adding a new tree to adjust the residual error between the prediction results of the previous tree and the true value, a new model was formed and used as the basis for the next model learning (J Li *et al* 2022). XGBoost increases the weight of training samples with high error rates and processes them multiple times with the aim of reducing the error rate (Kiangala and Wang 2021, Singha *et al* 2021). Therefore, this algorithm is insensitive to outliers and consistent against overfitting, which simplifies model selection (Shahhosseini *et al* 2019). For the mathematical details of the algorithm, see Chen and Guestrin 2016.

The ML library packages within Python, scikit-learn and XGBoost, were used to carry out the RF and XGBoost algorithms and CV. Each model was validated using a K–fold CV with 10 repeats. To conduct RF and XGBoost analysis, a grid search for model performance optimisation was carried out with the CV; the hyperparameter ranges and optimised values detected are shown in table 2.

2.3.3. Prediction performance assessment

Model performance was evaluated using the modified version of the Kling-Gupta Efficiency (KGEM) and its three components (equation 1): r represents the correlation coefficient between the simulated and observed time series; β (bias) is the ratio between the simulated and observed means (μ) (equation 2); and γ is the ratio of the coefficients of variation for both time series (equation 3). The optimal value of the KGEM and for each of the three components is 1. The KGEM indicator provides a useful assessment of model performance due to its decomposition into correlation (r), bias (β), and variability (γ). In this way, the model's ability to reproduce the temporal dynamics and distribution of nitrate concentration can be measured (Gupta *et al* 2009, Kling *et al* 2012).

$$KGEM = 1 - \sqrt{(r - 1)^2 + (\beta - 1)^2 + (\gamma - 1)^2} \tag{1}$$

$$\beta = \frac{\mu_{sim}}{\mu_{obs}} \tag{2}$$

$$\gamma = \frac{Coefficient\ \ of\ \ variation_{sim}}{Coefficient\ \ of\ \ variation_{obs}} \tag{3}$$

Table 2. Hyperparameter ranges and optimised values detected with grid search.

Algorithms	Parameter	Range	Optimum value
Random forest regression	n_estimators	100 to 1000	500
	max_depth	80, 90, 100, 110	110
	min_samples_leaf	2–10	3
	min_samples_split	2–12	10
	Bootstrap	True, False	False
XGBoost regressor	learning_rate	0.01, 0.05, 0.1, 0.2, 1	0.1
	max_depth	1–10	3
	Gamma	0–5	0
	min_child_weigh	1–10	4

Table 3. Variable importance information obtained after the analysis of mutual information and Pearson's coefficient and running the BorutaShap algorithm.

Features	Mutual Information	Pearson's Coefficient	BorutaShap	
			Mean Importance	Decision
Nitrogen (N)	1.18	0.94	4.53	Accepted
Piezometric level (PL)	0.87	0.48	−0.11	Accepted
Distance from river source (DRS)	0.81	0.58	−0.19	Accepted
Nitrates groundwater	0.68	0.71	−0.23	Rejected
Riparian forest quality (QBR)	0.68	0.50	−0.21	Rejected
Specific pollution sensitivity index (IPS)	0.64	0.53	−0.20	Rejected
Benthonic fauna of invertebrates (IBMWP index)	0.52	0.46	−0.17	Accepted
Nitrites (NO_2)	0.46	0.31	−0.21	Rejected
Total phosphorus (TP)	0.40	0.28	−0.09	Accepted
pH	0.30	−0.51	−0.10	Accepted
Ammonium (NH_4)	0.25	0.34	−0.19	Rejected
Streamflow	0.20	0.30	−0.18	Rejected
Water temperature (T_w)	0.18	0.07	−0.17	Rejected
Diffuse pollution (DP)	0.14	0.16	−0.19	Tentative
Precipitation	0.13	0.17	−0.22	Tentative
Air temperature (T_a)	0.13	−0.06	−0.18	Rejected
Dissolved oxygen (DO)	0.08	−0.26	−0.19	Rejected
Suspended solids (SS)	0.06	0.15	−0.23	Rejected
Biochemical oxygen demand over five days (BOD_5)	0.00	0.18	−0.23	Rejected

3. Results and discussion

3.1. Feature selection

Pearson's coefficient demonstrated the linear correlation between all candidate input parameters with the output parameter (table 3). The N ($r_p = 0.92$), nitrate-GW ($r_p = 0.70$), DRS ($r_p = 0.61$), and PL ($r_p = 0.58$) values showing higher Pearson correlation and the T_w ($r_p = -0.07$) and T_a ($r_p = 0.05$) values with the lowest Pearson correlation were identified as the most and the least influential parameters, respectively, when estimating the nitrate values. Regarding the predictor variables, a strong correlation of DRS was found with the PL (−0.95).

Table 3 shows the sensitivity analysis of applied MI for selecting dominant inputs. The highest MI scores were obtained with the N (1.15), PL (0.90), DRS (0.85), and nitrate-GW (0.68), and the lowest with DBO5 (0.00), SS (0.06), and DO (0.08). BorutaShap was applied to verify the Pearson and MI analysis, and the relative importance of features according to BorutaShap (table 3) indicated that N, DRS, piezometric level, IBMWP, TP and pH were the most important features for predicting nitrate concentration. The tentative features were DP and precipitation; the others were considered unimportant, and they should be omitted from the modelling process. The Pearson's coefficient, MI and BorutaShap values agreed on the three most influential parameters (N, PL and DRS), while the less influential parameters changed depending on the FS method.

The output of the FS methods was used to choose the input groups for the algorithms (table 4). Group 1 was composed of 19 features, and Group 2 of the 10 features with the highest value of the MI and Pearson correlation coefficient. Group 3 was similar to Group 2 but one variable (QBR) was excluded to increase the number of data; Group 4 was composed of the features selected using BorutaShap, and Group 5 was a mixture of the results found with MI, Pearson's coefficient (Group 3) and BorutaShap (Group 4). In Group 5, PL was excluded due to

Table 4. Input combinations based on Pearson's coefficient, mutual information and the BorutaShap algorithm to estimate nitrate concentration.

Models	Input Combinations	Input number	Data quantity
Group 1	N, PL, DRS, nitrate-GW, QBR, IPS, IBMWP, NO_2, TP, pH, NH_4, streamflow, T_W, DP, precipitation, T_a, DO, SS, BOD_5	19	240
Group 2	N, PL, DRS, nitrate-GW, QBR, IPS, IBMWP, NO_2, TP, pH	10	265
Group 3	N, PL, DRS, nitrate-GW, IPS, IBMWP, NO_2, TP, pH	9	420
Group 4	N, PL, DRS, IBMWP, TP, pH, DP, precipitation	8	427
Group 5	N, DRS, nitrate-GW, IPS, IBMWP, TP, pH, DP, precipitation	9	648

data availability and because it demonstrated a significant correlation with DRS, which could present collinearity. NO_2 was excluded due to data availability and because it did not show a strong relationship with nitrate concentration.

3.2. Modelling assessment

The nitrate concentration in the Júcar RB was predicted using RF and XGBoost algorithms with five groups of predictors. The KGEM indicator and its three parameters were calculated to evaluate the prediction accuracy of these models, and the values obtained in the validation stage are shown in figures 3(a) and (b). In all models with the RF algorithm the lineal correlation between simulated and observed is greater than 0.88, the bias was smaller than 10 percent and errors in the simulated variability less than 9 percent. The KGEM value ranged between 0.85 and 0.90, which means that there were no significant changes in the model's performance within the different groups. The difference in each of the parameters turned out to be only a few percent of the overall achievable range. However, a 4-percent increase in linear correlation was found with Group 3 when compared to Group 1. The best KGEM index was found within Group 5, which decreased the bias and increased lineal correlation. Meanwhile, the probability density function (PDF) of the residuals in validation shows (figure 3(c)) that all groups with RF algorithm were well-proportioned with lower mean and standard deviation values with high accumulation of errors in zero values. The differences observed between groups with the KGEM index are supported by the PDF.

In the models with XGBoost algorithm in the validation stage, the KGEM index range was between 0.77 and 0.87, the lineal correlation greater than 0.86, the bias smaller than 6 percent, and the variability smaller than 16 percent (figure 3(b)). In general, the XGBoost algorithm showed a systematic tendency to slightly underestimate the nitrate concentration in the validation. Group 5 showed the best result, decreasing the bias in simulated to 4 percent (figure 3(b)), and improving the model performance by 2 percent compared with Group 1. The PDF shows that the errors of Group 5 were well-proportioned with lower mean and high accumulation in zero values, whereas the other groups showed a higher standard deviation of errors (figure 3(d)). However, after using CV, the predictive performance of the models with XGBoost improved and reached a behavior similar to RF (figure 3(f)).

Group 5, which consisted of the variables with the best MI and BoruraShap scores, was identified as the optimal input combination for the two algorithms. It provided high lineal correlation, was unbiased (RF) or slightly biased (XGBoost), and the variability was smaller. Moreover, mean and standard deviation of errors had high accumulation in zero values. Likewise, the weakest performances in the validation with the two algorithms were related to Group 2, which consisted of the 10 variables with the best MI scores. It demonstrated high lineal correlation, and small bias; however, errors in the simulated variability are widespread (38 percent). After applying CV, Groups 1, 2, 3 and 4 displayed a similar behavior (figures 3(e) and (f)), and Group 5 still produced the best performance.

The plots simulated and observed nitrate values are shown in figure 4, comparing the performance of the two predictive algorithms applying CV with Group 5. The models showed a pattern of nitrate distribution along the river similar to the observed data, with differences existing mainly downstream of the watershed, where the models slightly underestimated the nitrate concentration (figures 4(a) and (b)). In general, the probability of identifying high nitrate concentrations increased in the middle and downstream of the watershed. The models fit the temporal variability of nitrate concentrations along the river. There had been a slight decrease in recent years, and this behavior is represented in the models. Moreover, the seasonal variability was in accordance with the observed values, with nitrate concentration higher in autumn and winter, and decreasing in summer. However, there was a slight underfitting in the values simulated in autumn and winter with the two algorithms downstream of the basin (figures 4(c) and (d)).

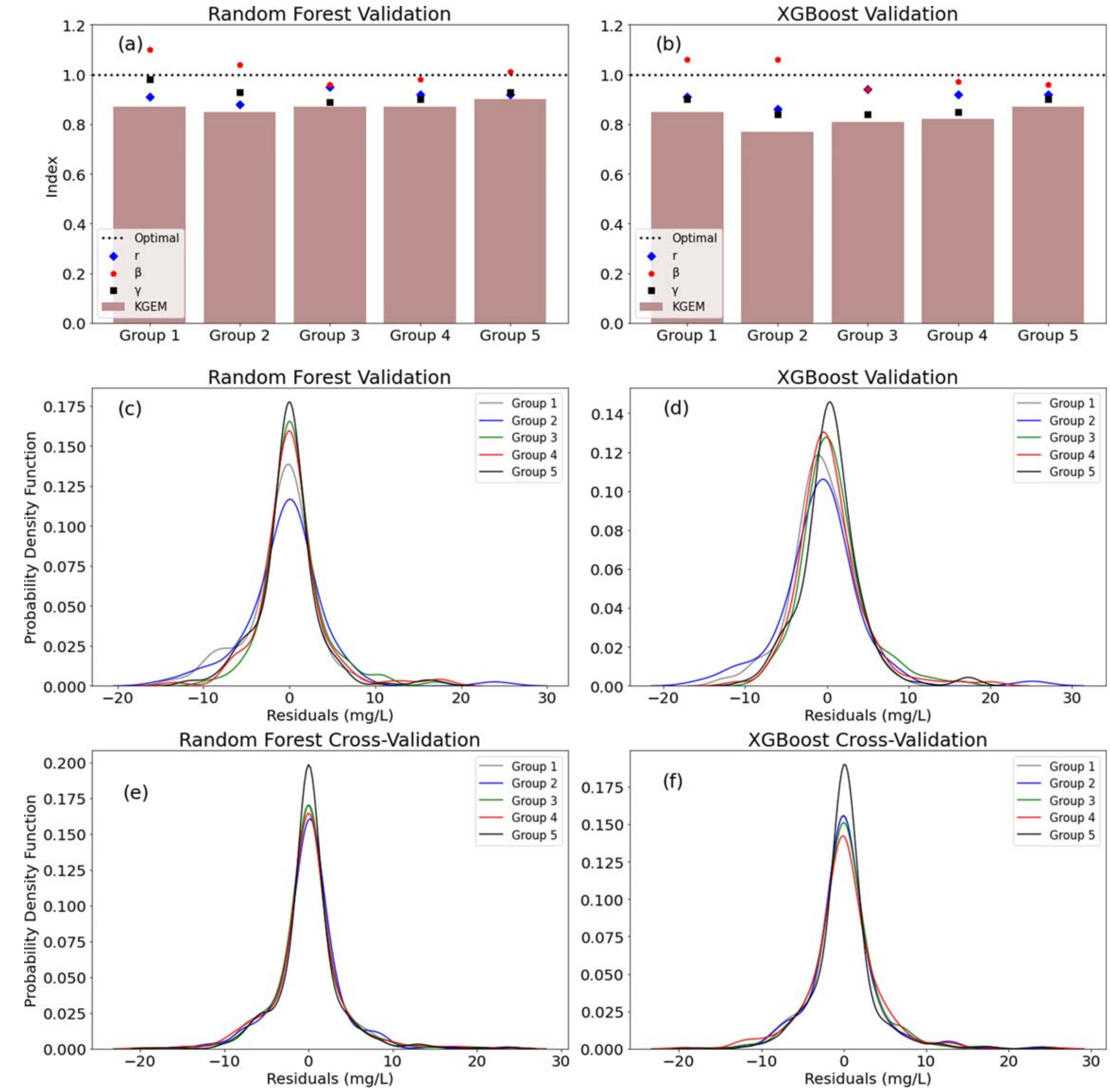

Figure 3. The KGEM components for assessing the prediction of nitrate concentration using RF (a) and XGBoost (b) algorithms. Probability density function of residuals for the prediction of nitrate concentration using RF (c) and XGBoost (d) algorithms in the validation stage and after apply K–fold cross validation with 10 repeats (e and f). r = correlation coefficient; β = bias ratio; γ = ratio of the coefficients of variation; KGEM = Modified Kling-Gupta Efficiency.

3.3. Importance of conditioning factors

The importance of the driving features in the modelling process is shown in figure 5. N was the most important feature in the prediction of nitrate concentration using RF and XGBoost algorithms. This result agrees with the MI, Pearson's coefficient and BorutaShap; however, there were differences between groups and algorithms in terms of ranking the features. Most important among the other features for the prediction of nitrate concentration are the following: DRS in Group 5 with RF, and Groups 4 and 5 with XGBoost; nitrate-GW in Groups 3 and 5 with both algorithms; precipitation in Groups 4 and 5 with both algorithms; PL in all groups in both algorithms (with the exception of Group 5); and pH and total P in all groups in both algorithms. Group 2, which performed with less accuracy using the two algorithms, gave a high importance (88 percent RF—90 percent XGBoost) to N, while in Group 5 with RF, the importance of N is 57 percent. Of all the variables used in the prediction of nitrate concentration, the least contributing variables were NO_2, NH_4, DO, SS, BOD_5, Tw, Ta, streamflow and QBR.

4. Discussion

4.1. Comparison of models and feature selection approaches

The models used with the RF and XGBoost algorithms are reliable when estimating the nitrate concentration in the Júcar RB. However, the difference in the calculation procedures of feature selection methods and algorithms

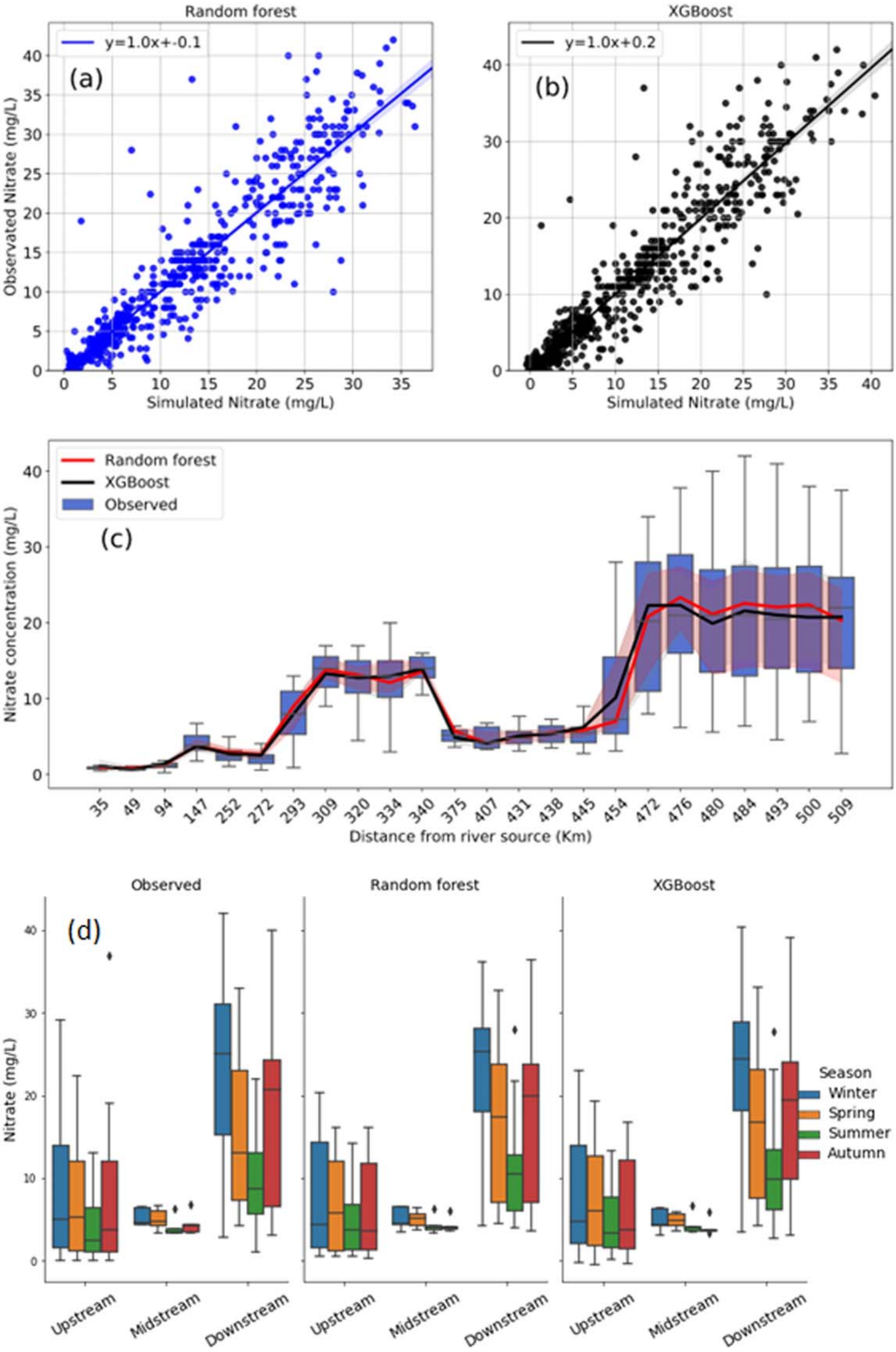

Figure 4. Linear regression for simulated and observed values of Group 5 (best model) applying RF (a) and XGBoost (b) algorithms. (c) Monthly nitrate mean concentration observed (box squares without including outliers), simulated (continuous line, first and third lower and upper shaded quartiles, respectively) and (d) seasonal nitrate concentration along the Júcar river Basin applying RF and XGBoost algorithms.

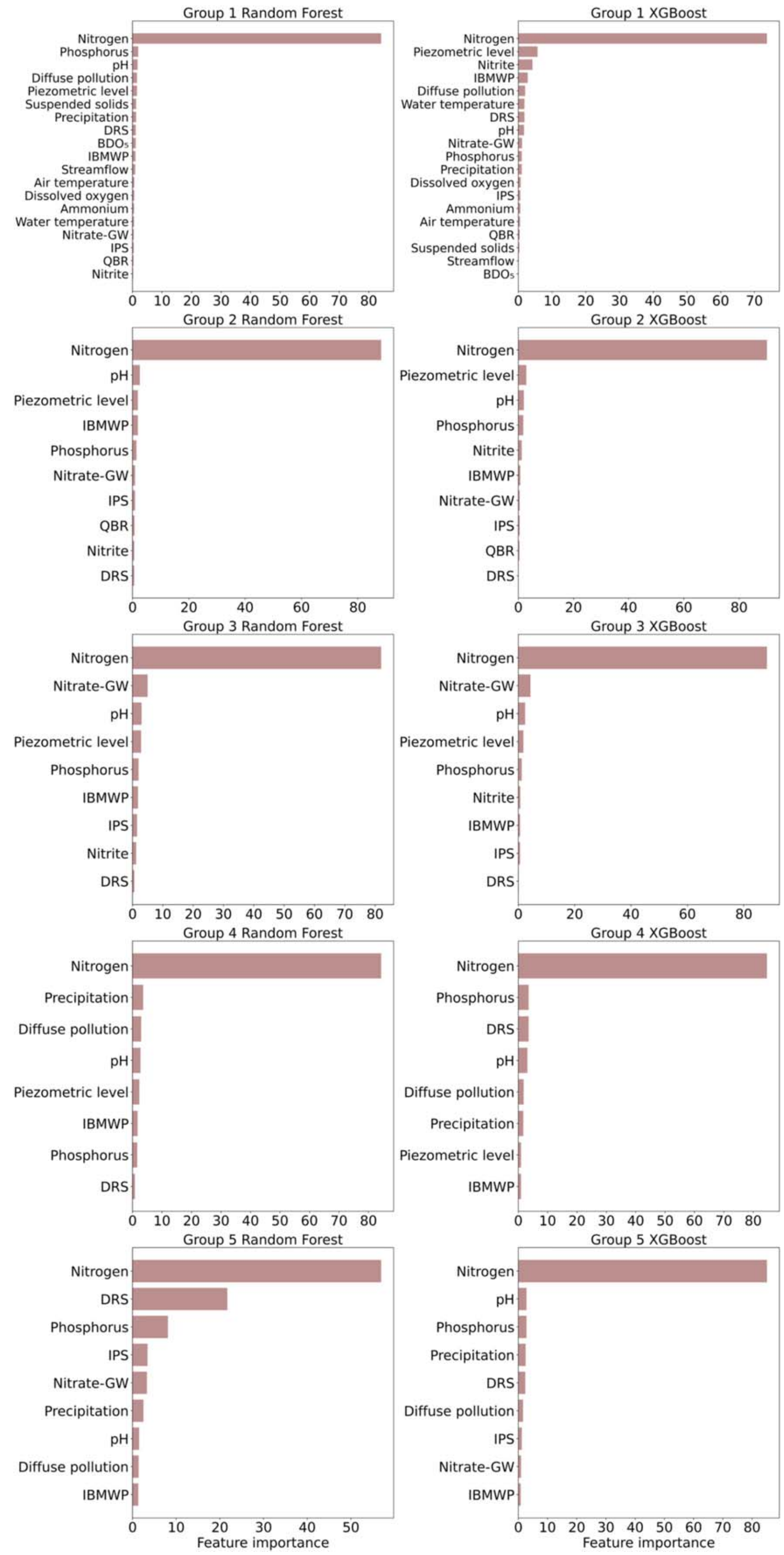

Figure 5. Feature importance results of the RF and XGBoost algorithms with different groups of variables.

resulted in different model performances. Models in Group 2, consisting of the 10 features with the best MI score, performed the worst with the two algorithms. This could be because MI assesses the features independently without considering their context, and the features were selected in a univariate way. Therefore, MI was not able to deal directly with the problem of redundant inputs (Nourani *et al* 2017, Effrosynidis and Arampatzis 2021). Models in Group 3 were similar to Group 2, but the removal of a feature with only few data improved the model performance by 2 percent. Models in Group 4, consisting of the BorutaShap results (selection of 8 out of 19 features) improved the model performance from Group 2 by 2 percent. Although BorutaShap is a new algorithm, it has recently been used in different fields, performing well in terms of feature reduction and predictive accuracy (Kleiman *et al* 2021, Ghosh and Chaudhuri 2022, Peiró-Signes *et al* 2022). It reduces the number of features by including only the relevant ones without compromising the model performance, and the Shap value embedded in the algorithm adds an important explanatory capacity that reduces the overfitting problem (Ghosh and Chaudhuri 2022).

The highest performance was found with Group 5 (merging Groups 3 and 4) with the two algorithms. Combining the results of the two selection methods and knowledge of the data allowed variables that were highly correlated and those that provided few data to be excluded. PL depends on DRS, and removing PL from the predictors reduced the model complexity and the cost of prediction and increased the sample size of the dataset. Sample size had a significant impact on modelling and prediction performance in this study, and the increase of training data and smaller set of features decreased the variance among the residuals. In this way, the performance of the model was improved. Similar results were found by Shahhosseini *et al* 2019, Zamani Joharestani *et al* 2019 and Effrosynidis and Arampatzis 2021.

Comparing the two algorithms in the validation stage for Group 5, the RF resulted in a slightly better performance (3 percent) in respect to bias and variance. However, after applying CV the performance of XGBoost improved (4 percent), while RF remained the same. Therefore, either algorithm could be used for nitrate prediction, as the difference between the two algorithms was 1 percent. The improvement with CV for the XGBoost algorithm was possibly due to the fact that successive trees gave extra weight to points incorrectly predicted during the previous analysis and finally a weighted vote was taken for the prediction (Fan *et al* 2018). After using CV, both models were able to recognise the complex interactions between conditioning factors, and Tomperi *et al* 2017 reported an increase in the accuracy of the prediction of AI models after applying the CV method.

On the other hand, the results revealed how sensitive XGBoost is to the wrong features being selected. In Groups 2, 3, and 4, the XGBoost metrics decreased for the validation dataset. In contrast, RF showed a more robust model, and introducing wrong features to RF did not change the model performance considerably, as it maintained a similar performance level. In other research using RF and XGBoost algorithms, the authors reported that they obtained the best performance with XGBoost, although the difference with RF was small (Fan *et al* 2018, Zhong *et al* 2019, Kiangala and Wang 2021, Peiró-Signes *et al* 2022). XGBoost and RF are ensemble algorithms; therefore, it is difficult to explain their predictions, and each one has different limitations. The performance of RF depends on the amount of data used in the training dataset (Ghimire *et al* 2022), while XGBoost presents less accurate results when dealing with imbalanced data (Kiangala and Wang 2021).

The models used with XGBoost and RF algorithms are substantially higher than the traditional hydrological models applied in the Júcar RB. The coupling of hydrological and water quality models in the Júcar RB found 58 percent of lineal correlation, a bias smaller than 20 percent, and the variability was 25 percent (Dorado-Guerra *et al* 2021). ML algorithms improved the correlation, bias and variability measures reached with the coupling of hydrological models in the Júcar RB by 40 and 37 percent with RF and XGBoost algorithms, respectively, with the lineal correlation the parameter that improved the most. Similar results were found by Wu *et al* 2017, who reported that AI algorithms are statistically better than hydrologic models.

4.2. Use of extrinsic features of surface water bodies and their effect on nitrate pollution

It can be inferred that it is possible to model the nitrate concentration in SW in the Júcar RB using N, DRS, P, IPS pH, nitrate-GW, precipitation, DP and IBMWP, the features representative of weather, location, ecological status, water quality and anthropogenic effects. This approach could be considered as a methodology to predict nitrate concentration, especially in data-scarce areas, but it must be validated in the other catchments of the region. Other studies showed that location and precipitation were important driving factors affecting water quality in rivers and aquifers (Ha *et al* 2020, He *et al* 2022, Wang *et al* 2022).

The results show that the high nitrate concentration in the Júcar RB is linked to high nitrogen zones (figure 6), and that the relationship between these two variables is lineal as shown by Pearson's correlation. Other studies showed similar results, in which nitrogen was the main predictor of nitrates (Oehler and Elliott 2011). Nitrogen leaches when transformed into nitrate form, and the main issue then with nitrates is their mobility in the soil and the fact that they can persist in SW and GW (Defterdarović *et al* 2021). Agricultural activity is the main source of nitrogen in the watershed (Dorado-Guerra *et al* 2021); therefore, DP is the most probable cause for the higher nitrate probabilities and the increase of the nitrate concentrations in the river.

The DRS exhibited a positive effect on nitrate concentration in SW in the Júcar river (figure 6), and a similar result in GW was found by Rodriguez-Galiano *et al* 2014 and He *et al* 2022. This may be because the nitrate

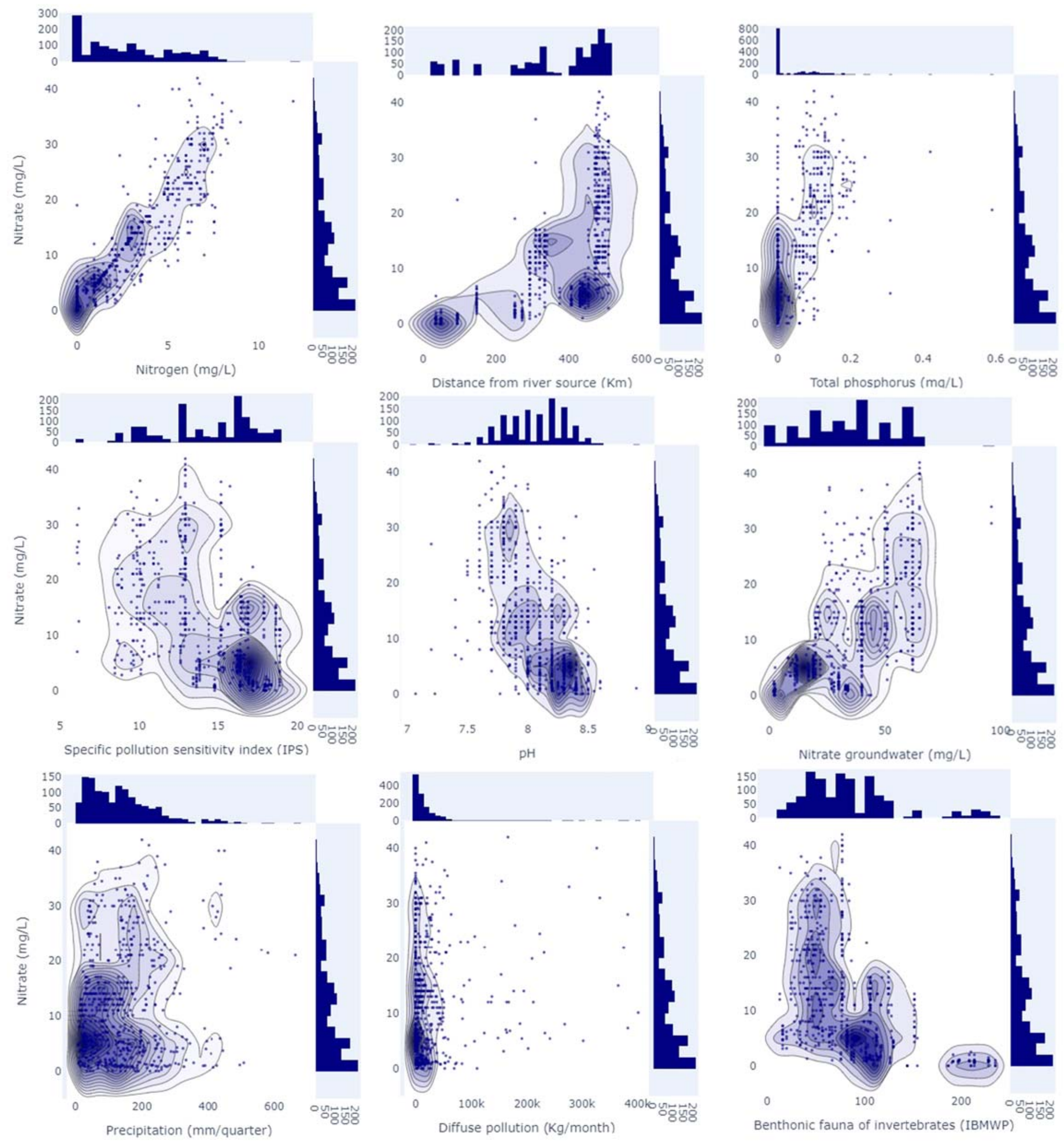

Figure 6. Scatter plot overlaid on a density contour plot to show a correlation between predictor features of group 5 with RF algorithm and the target variable.

pollution is associated with agricultural zones located in the downstream of the watershed, while in the upstream the land use is forest (Dorado-Guerra *et al* 2021). Therefore, DRS contributed significant information to help identify polluted areas.

Precipitation was the most influential meteorological variable with relative importance, though a weak positive effect of precipitation on SW nitrate was detected by the two algorithms used. In this study, precipitation above 500 mm/trimester was associated with high nitrate concentration in SW (figure 6); as nitrate inputs were mainly from diffuse sources, rise of nitrate concentration takes place mainly in winter and spring when precipitation is high (figures 4(c) and (d)). However, the influence of precipitation on the SW nitrate concentration is complex, as shown in figure 6. For example, high rainfall increases the streamflow resulting in the dilution of SW chemical components (Romero *et al* 2007; Temino-Boes *et al* 2021), which can also promote crops to uptake nitrogen (Sieling and Kage 2006). The precipitation would then have positive and negative effects on nitrate concentration in SW.

TP was another important factor for predicting nitrates in SW, with similar results found by Oehler and Elliott 2011. TP above 0.1 mg l^{-1} was associated with high nitrate concentration in the Júcar RB (figure 6), which might be an indicator of the N:P ratio controlling important N speciation processes through temporary plant uptake and decay (Ensign and Doyle 2006). As for pH, there was a negative relationship with nitrate concentration, perhaps due to the fact that increasing pH affects microbial activity and decreases the nitrification process (Chen *et al* 2006), and pH levels above 8.25 and below 7.4 were associated with the lowest nitrate concentration (figure 6). The relationship between the high nitrate concentration in GW was not clearly related to the high nitrate levels in SW, because this relationship depends on the river-aquifer interaction. However, in a previous study, it was shown that there is a high linear correlation between nitrate content in both GW and SW when river and aquifer are connected (Dorado-Guerra *et al* 2021).

Nitrate is an important predictor of diatoms index IPS and macroinvertebrates index IBMWP (Valerio *et al* 2021), and several studies have shown that diatom distribution is highly dependent on nitrates, which have fast growth rates that allow them to react faster to chemical changes and detect the first step of degradation (Doung *et al* 2007, Tan *et al* 2017, Karaouzas *et al* 2019). The relationship between nitrate and IPS and IBMWP indices was negative in this study (figure 6); IPS values above 16 were related with the lowest nitrate concentration, while IBMWP values above 80 were related with the lowest nitrate concentration (figure 6).

5. Conclusions

This paper explores the potential of feature selection and artificial intelligence algorithms to model nitrate concentration in surface water bodies in areas with water scarcity and high interaction between rivers and aquifers. RF and XGBoost successfully modelled the nitrate concentration in the Júcar RB and enabled recognition of the complex interactions between conditioning factors. FS methods are useful tools, but they need to be combined with local knowledge of the dataset, as the amount of data available and high correlation between predictor features affect the performance of the models. Nitrogen, total phosphorus and location were the strongest predictor factors for nitrate concentration in surface water bodies in the Júcar RB, because they accounted for approximately 88 percent of the nitrate variation. On the other hand, RF and XGBoost models obtained better performance than hydrological models in the prediction of nitrate concentration in surface water bodies of Júcar RB.

Acknowledgments

We appreciate the help provided by the Júcar River Basin District Authority (CHJ), who gathered field data. The first author's research is partially funded by a PhD scholarship from the food research stream of the program 'Colombia Científica—Pasaporte a la Ciencia', granted by the Colombian Institute for Educational Technical Studies Abroad (Instituto Colombiano de Crédito Educativo y Estudios Técnicos en el Exterior, ICETEX). The authors thank the Spanish Research Agency (AEI) for the financial support to RESPHIRA project (PID2019-106322RB-100)/AEI/10.13039/501100011033.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://doi.org/https://aps.chj.es/siajucar/.

Conflict of interest

The authors declare no conflicts of interest.

ORCID iDs

Diana Yaritza Dorado-Guerra https://orcid.org/0000-0001-8662-0160
Gerald Corzo-Pérez https://orcid.org/0000-0002-2773-7817
Javier Paredes-Arquiola https://orcid.org/0000-0003-3198-2169
Miguel Ángel Pérez-Martín https://orcid.org/0000-0002-4733-0862

References

Alba-Tercedor J *et al* 2002 Caracterización del estado ecológico de ríos mediterráneos ibéricos mediante el índice IBMWP (antes BMWP') *Limnetica* **21** 175–85

Alcon F, Zabala A and Martínez-Paz J 2022 Assessment of social demand heterogeneity to inform agricultural diffuse pollution mitigation policies *Ecol. Econ.* **191**

Bagherzadeh F, Mehrani M, Basirifard M and Roostaei J 2021 Journal of water process engineering comparative study on total nitrogen prediction in wastewater treatment plant and effect of various feature selection methods on machine learning algorithms performance . *J. Water Process Eng.* **41** 102033

Bao Q, An D, Thang N, Reza A and Islam T 2022 Random forest and nature-inspired algorithms for mapping groundwater nitrate concentration in a coastal multi-layer aquifer system *J. Clean. Prod.* **343** 130900

Bouraoui F and Grizzetti B 2014 Modelling mitigation options to reduce diffuse nitrogen water pollution from agriculture *Sci. Total Environ.* **468–469** 1267–77

Breiman L 2001 Random forests *Mach. Learn.* **45** 5–32

Castrillo M and López A 2020 Estimation of high frequency nutrient concentrations from water quality surrogates using machine learning methods *Water Res.* **172**

Cemagref 1982 Etude des méthodes biologiques d'appréciation quantitative de la qualité des eaux. rapport qe lyon & mdash agence de l'eau rhone-méditerranée- corse

Chen S, Ling J and Blancheton J 2006 Nitrification kinetics of biofilm as affected by water quality factors *Aquac. Eng.* **34** 179–97

Chen T and Guestrin C 2016 XGBoost: a Scalable tree boosting system *Preprints, the 22nd ACM SIGKDD Int. Conf. 19*

Confederación Hidrográfica del Júcar 2022a Plan hidrológico de la demarcación hidrográfica del Júcar. Memoria -anejo 12. Evaluación del estado de las masas de agua superficial y subterránea. Ciclo de planificación hidrológica 2022-2027.(https://chj.es/Descargas/ProyectosOPH/Consulta%20publica/PHC-2021-2027/PHJ/PHJ2227_CP_Anejo12_Estado.pdf) (accessed on 10 March 2022)

Confederación Hidrográfica del Júcar 2022b Plan hidrológico de la demarcación hidrográfica del Júcar Memoria. Ciclo de planificación hidrológica 2022-2027. (https://chj.es/es-es/medioambiente/planificacionhidrologica/Documents/Plan-Hidrologico-cuenca-2021-2027/PHC/Documentos/PHJ2227_Memoria_20220329.pdf) (accessed on 12 May 2022)

Confederación Hidrográfica del Júcar 2022c Plan hidrológico de la demarcación hidrográfica del Júcar. Memoria -anejo 7. Evaluación de las presiones, impacto y riesgo de las masas de agua. Ciclo de planificación hidrológica 2022-2027. (https://chj.es/es-es/medioambiente/planificacionhidrologica/Documents/Plan-Hidrologico-cuenca-2021-2027/PHC/Secretaria%20General%20Tecnica/PHJ2227_SGT_Anejo07_InvPresiones.pdf) (accessed on 22 September 2022)

Cover T M and Thomas J A 2006 Elements of information theory second edition solutions to problems. (https://cpb-us-w2.wpmucdn.com/sites.gatech.edu/dist/c/565/files/2017/01/solutions2.pdf). (accessed on 15 October 2021)

Defterdarović J *et al* 2021 Determination of soil hydraulic parameters and evaluation of water dynamics and nitrate leaching in the unsaturated layered zone: a modeling case study in central croatia *Sustain.* **13** 1–20

Dorado-Guerra D Y, Corzo-Perez G, Paredes-Arquiola J and Perez-Martin M A 2022 Dataset on surface water features of the Júcar River Basin Valencia (Spain) to Predict Nitrate Concentration. 4TU. ResearchData. Dataset.

Dorado-Guerra D Y, Paredes-Arquiola J, Pérez-Martín M Á and Hermann H T 2021 Integrated surface-groundwater modelling of nitrate concentration in mediterranean rivers, the júcar river basin district, Spain *Sustain.* **13**

Doung T, Feurtet-Mazel A, Coste M, Dam K and Boudou A 2007 Dynamics of diatom colonization process in some rivers influenced by urban pollution (Hanoi , Vietnam) *Ecol. Indic.* **7** 839–51

Effrosynidis D and Arampatzis A 2021 An evaluation of feature selection methods for environmental data *Ecol. Inform.* **61** 101224

Ensign S H and Doyle M W 2006 Nutrient spiraling in streams and river networks *J. Geophysical Research: Biogeosciences* **111** 1–13

European Environmental Agency (EEA) 2021 CorineLand Cover. 2021. Available online: https://eea.europa.eu/publications/COR0-landcover (accessed on 30 March2021)

Fan J, Yue W, Wu L, Zhang F, Cai H, Wang X, Lu X and Xiang Y 2018 Evaluation of SVM, ELM and four tree-based ensemble models for predicting daily reference evapotranspiration using limited meteorological data in different climates of China *Agric. For. Meteorol.* **263** 225–41

Garabaghi F H 2022 Performance Evaluation of Machine Learning Models with Ensemble Learning Approach in Classification of Water Quality Indices Based on Different Subset of Features *Research Square* 1–36

Gervasi O, Murgante B, Misra S, Maria A and Goos G 2022 *Computational Science and Its Applications—ICCSA* 13379, 1–733

Ghimire S, Deo R C, Casillas-Pérez D and Salcedo-Sanz S 2022 Boosting solar radiation predictions with global climate models, observational predictors and hybrid deep-machine learning algorithms *Appl. Energy* **316** 119063

Ghosh I and Chaudhuri T D 2022 Integrating navier–stokes equation and neoteric iforest-borutashap-facebook prophet framework for stock market prediction: an application in indian context *Expert Syst. Appl.* **210**

Grinsven H J M V, Bouwman L, Cassman K G, Es H M, Van, Mccrackin M L and Beusen A H W 2015 Losses of ammonia and nitrate from agriculture and their effect on nitrogen recovery in the european union and the united states between 1900 and 2050 *J. Environ. Qual.* **44** 356–67

Grizzetti B, Vigiak O, Udias A, Aloe A, Zanni M, Bouraoui F and Pistocchi A 2021 How EU policies could reduce nutrient pollution in European inland and coastal waters *Glob. Environ. Chang.* **69** 102281

Gu X, Sun H, Tick G R, Lu Y, Zhang Y, Zhang Y and Schilling K 2020 Identification and scaling behavior assessment of the dominant hydrological factors of nitrate concentrations in streamflow *J. Hydrol. Eng.* **25** 06020002

Gupta H V, Kling H, Yilmaz K K and Martinez G F 2009 Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling *J. Hydrol.* **377** 80–91

Ha N, Nguyen H Q and Cung N 2020 Estimation of nitrogen and phosphorus concentrations from water quality surrogates using machine learning in the Tri An Reservoir, Vietnam *Environ Monit Assess* **192**

He S, Wu J, Wang D and He X 2022 Predictive modeling of groundwater nitrate pollution and evaluating its main impact factors using random forest *Chemosphere* **290** 133388

IGME-DGA 2012 Trabajos de la Actividad 4 'Identificación y caracterización de la interrelación que se presenta entre aguas subterráneas, cursos fluviales, descargas por manantiales, zonas húmedas y otros ecosistemas naturales de especial interés hídrico DHJ. Institut. 1-141 (https://chj.es/Descargas/ProyectosOPH/Consulta%20publica/PHC-2015-2021/ReferenciasBibliograficas/AguasSubterraneas/IGME-DGA,2009.Act04_RelacSuperf_SubtMEMORIA%20RESUMEN.pdf) (accessed on 20 October 2021)

Izzuan H, Yusri H, Afhzan A, Rahim A, Lailatul S, Hassan M, Shairah I, Halim A and Abdullah N E 2022 Water Quality Classification Using SVM And XGBoost Method. IEEE 13th Control *Syst. Grad. Res. Colloq.* 231–6

Jamei M, Karbasi M and Malik A 2022 Developing hybrid data-intelligent method using Boruta-random forest optimizer for simulation of nitrate distribution pattern *Agricultural Water Management* **270** 107715

Karaouzas I, Smeti E, Kalogianni E and Skoulikidis N T 2019 Ecological status monitoring and assessment in Greek rivers : Do macroinvertebrate and diatom indices indicate same responses to anthropogenic pressures ? *Ecol. Indic.* **101** 126–32

Keany E 2020 BorutaShap : A wrapper feature selection method which combines the Boruta feature selection algorithm with Shapley values (https://doi.org/10.5281/ZENODO.4247618) (https://zenodo.org/record/4247618#.Y6HbjnbMLIU) (accessed on 9 May 2022)

Keany E 2021 BorutaShap 1.0.16 [WWW Document]. URL (https://pypi.org/project/BorutaShap/) (accessed 8.5.22)

Kiangala S K and Wang Z 2021 An effective adaptive customization framework for small manufacturing plants using extreme gradient boosting-XGBoost and random forest ensemble learning algorithms in an Industry 4.0 environment *Mach. Learn. with Appl.* **4** 100024

Kleiman M, Barenholtz E and Galvin J 2021 Screening for early-stage alzheimer's disease using optimized feature sets and machine learning *HHS Public Access* **81** 355–66

Kling H, Fuchs M and Paulin M 2012 Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios *J. Hydrol.* **424–425** 264–77

Kursa M B and Rudnicki W R 2010 Feature selection with the boruta package *J. Stat. Softw.* **36** 1–13

Latif S D, Azmi M S B N, Ahmed A N, Fai C M and El-Shafie A 2020 Application of artificial neural network for forecasting nitrate concentration as a water quality parameter: a case study of feitsui reservoir *Taiwan. Int. J. Des. Nat. Ecodynamics* **15** 647–52

Li J, An X, Li Q, Wang C, Yu H, Zhou X and Geng Y 2022 Application of XGBoost algorithm in the optimization of pollutant concentration *Atmos. Res.* **276** 106238

Li L, Qiao J, Yu G, Wang L, Li H, Liao C and Zhu Z 2022 Interpretable tree-based ensemble model for predicting beach water quality *Water Res.* **211** 118078

Ma M, Zhao G, He B, Li Q, Dong H, Wang S and Wang Z 2021 XGBoost-based method for flash flood risk assessment *J. Hydrol.* **598** 126382

Ministerio para la transción ecológica y el reto Demográfico 2020 MInforme de seguimiento de la directiva 91/676/CEE contaminación del agua por nitratos utilizados en la agricultura (https://miteco.gob.es/es/agua/temas/estado-y-calidad-de-las-aguas/informe-2016-2019_tcm30-518402.pdf) (accessedon 30 April 2022)

Munné A, Prat N, Solà C, Bonada N and Rieradevall M 2003 A simple field method for assessing the ecological quality of riparian habitat in rivers and streams : QBR index *Aquat. Conserv Mar. Freshw. Ecosyst* **163** 147–63

Nasir N, Kansal A, Alshaltone O, Barneih F, Sameer M, Shanableh A and Al-shamma A 2022 Journal of water process engineering water quality classification using machine learning algorithms *J. Water Process Eng.* **48** 102920

Nourani V, Andalib G and Dąbrowska D 2017 Conjunction of wavelet transform and SOM-mutual information data pre-processing approach for AI-based Multi-Station nitrate modeling of watersheds *J. Hydrol.* **548** 170–83

Oehler F and Elliott A H 2011 Science of the total environment predicting stream n and p concentrations from loads and catchment characteristics at regional scale : a concentration ratio method *Sci. Total Environ.* **409** 5392–402

Ortega-Reig M, Sanchis-Ibor C, Palau-Salvador G, García-Mollá M and Avellá-Reus L 2017 Institutional and management implications of drip irrigation introduction in collective irrigation systems in Spain *Agric. Water Manag.* **187** 164–72

Pang S, Wang X, Melching C S, Guo H and Li W 2022 Identification of multilevel priority management areas for diffuse pollutants based on streamflow continuity in a water-deficient watershed *J. Clean. Prod.* **351** 131322

Paredes-Arquiola J 2021 Manual técnico del modelo respuesta rápida del estado ambiental (R2EA) de masas de agua superficiales continentales Universitat Politècnica de València. (https://aquatool.webs.upv.es/files/manuales/rrea/ManualT%C3%A9cnicoModeloRREA_V3.pdf) (accessed on 20 November 2021)

Peiró-Signes Á, Segarra-Oña M, Trull-Domínguez Ó and Sánchez-Planelles J 2022 Exposing the ideal combination of endogenous–exogenous drivers for companies' ecoinnovative orientation: Results from machine-learning methods *Socioecon. Plann. Sci.* **79**

Peral García C, Navascués Fernández-Victorio B and Ramos Calzado P 2021 Serie de precipitación diaria en rejilla con fines climáticos Ser. Precipitación Diaria en Rejilla Con Fines Climáticos. (https://aemet.es/documentos/es/conocermas/recursos_en_linea/publicaciones_y_estudios/publicaciones/NT_24_AEMET/NT_24_AEMET.pdf) (accessed on 14 July 2021), pages 1-30

Pérez-Martín M A, Estrela T, Andreu J and Ferrer J 2014 Modeling water resources and river-aquifer interaction in the Júcar River Basin, Spain *Water Resour. Manag.* **28** 4337–58

Poikane S, Kelly M G, Salas F, Pitt J, Jarvie H P, Claussen U, Leujak W, Lyche A, Teixeira H and Phillips G 2019 Nutrient criteria for surface waters under the european water frame- work directive : current state-of-the-art , challenges and future outlook *Sci. Total Environ.* **695**

Rafiei V, Nejadhashemi A P, Mushtaq S, Bailey R T and An-vo D 2022 Groundwater-surface water interactions at wetland interface : Advancement in catchment system modeling *Environ. Model. Softw.* **152** 105407

Rodriguez-Galiano V, Mendes M P, Garcia-Soldado M J, Chica-Olmo M and Ribeiro L 2014 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.* **476–477** 189–206

Rodriguez-Galiano V F, Luque-espinar J A, Chica-olmo M and Mendes M P 2018 Feature selection approaches for predictive modelling of groundwater nitrate pollution : An evaluation of fi lters, embedded and wrapper methods *Sci. Total Environ.* **624** 661–72

Romero I, Moragues M, González del Río J, Hermosilla Z, Sánchez-Arcilla A, Sierra J P and Mösso C 2007 Nutrient behavior in the júcar estuary and plume *J. Coast. Res.* **10047** 48–55

Shahhosseini M, Martinez-Feria R A, Hu G and Archontoulis S V 2019 Maize yield and nitrate loss prediction with machine learning algorithms *Environ. Res. Lett.* **14**

Sieling K and Kage H 2006 N balance as an indicator of N leaching in an oilseed rape—winter wheat—winter barley rotation *Agriculture, Ecosystems & Environment* **115** 261–9

Singh B and Craswell E 2021 Fertilizers and nitrate pollution of surface and ground water : an increasingly pervasive global problem *SN Appl. Sci.* **3** 1–24

Singh S, Anil A G, Kumar V, Kapoor D, Subramanian S, Singh J and Ramamurthy P C 2022 Nitrates in the environment : a critical review of their distribution, sensing techniques, ecological effects and remediation *Chemosphere* **287** 131996

Singha S, Pasupuleti S, Singha S S, Singh R and Kumar S 2021 Prediction of groundwater quality using efficient machine learning technique *Chemosphere* **276** 130265

Tyralis H, Papacharalampous G and Langousis A 2019 A brief review of random forests for water scientists and practitioners and their recent history inwater resources *Water* **2019** 910

Tan X, Zhang Q, Burford M A, Sheldon F and Bunn S E 2017 Benthic diatom based indices for water quality assessment in two subtropical streams. front. microbiol. 8601file///c/users/a315-21-99m2/documents/articulo artif *Intell. Intell.* **8**

Temino-Boes R, García-Bartual R, Romero I and Romero-Lopez R 2021 Future trends of dissolved inorganic nitrogen concentrations in Northwestern Mediterranean coastal waters under climate change *J. Environ. Manage.* **282** 111739

Thornhill I, Ho J G, Zhang Y, Li H, Ho K C, Miguel-Chinchilla L and Loiselle S A 2017 Prioritising local action for water quality improvement using citizen science; a study across three major metropolitan areas of China *Sci. Total Environ.* **584–585** 1268–81

Tomperi J, Koivuranta E and Leiviskä K 2017 Journal of water process engineering predicting the effluent quality of an industrial wastewater treatment plant by way of optical monitoring . *J. Water Process Eng.* **16** 283–9

Tzilivakis J, Warner D J, Green A and Lewis K A 2021 A broad-scale spatial analysis of the environmental benefits of fertiliser closed periods implemented under the Nitrates Directive in Europe *J. Environ. Manage.* **299** 113674

Valerio C, Stefano L, De, Martínez-muñoz G and Garrido A 2021 Science of the total environment a machine learning model to assess the ecosystem response to water policy measures in the Tagus River Basin (Spain) *Sci. Total Environ.* **750** 141252

Vergara J R and Estévez P A 2014 A review of feature selection methods based on mutual information *Neural Comput. Appl.* **24** 175–86

Wang X, Liu X, Wang L, Yang J, Wan X and Liang T 2022 A holistic assessment of spatiotemporal variation, driving factors, and risks influencing river water quality in the northeastern Qinghai-Tibet Plateau *Sci. Total Environ.* **851**

Wu R, Painumkal J T, Volk J M and Liu S 2017 *Parameter Estimation of Nonlinear Nitrate Prediction Model Using Genetic Algorithm* 1893–9

Yuan L, Sinshaw T and Forshay K J 2020 Review of watershed-scale water quality and nonpoint source pollution models *Geosci.* **1** 1–33

Zamani Joharestani M, Cao C, Ni X, Bashir B and Talebiesfandarani S 2019 PM2.5 prediction based on random forest, XGBoost, and Deep Learning Using Multisource Remote Sensing Data *Atmosphere* **10** 373

Zhong L, Hu L and Zhou H 2019 Deep learning based multi-temporal crop classification *Remote Sens. Environ.* **221** 430–43

Zhu M, Wang J, Yang X, Zhang Y, Zhang L, Ren H, Wu B and Ye L 2022 A review of the application of machine learning in water quality evaluation *Eco-Environment Heal.* **1** 107–16