



Spatial prediction of groundwater nitrate through machine learning and stakeholder co-creation: the Chalk aquifer case in East Anglia, UK

V. Gómez-Escalonilla , H. Baron, T. Read , J. Watson , M. Rodríguez Del Rosario , G. Darch , V. Keller , N. Rickards & P. Martínez-Santos

To cite this article: V. Gómez-Escalonilla , H. Baron, T. Read , J. Watson , M. Rodríguez Del Rosario , G. Darch , V. Keller , N. Rickards & P. Martínez-Santos (18 Sep 2026): Spatial prediction of groundwater nitrate through machine learning and stakeholder co-creation: the Chalk aquifer case in East Anglia, UK, International Journal of River Basin Management, DOI: [10.1080/15715124.2026.2706022](https://doi.org/10.1080/15715124.2026.2706022)

To link to this article: <https://doi.org/10.1080/15715124.2026.2706022>



© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



[View supplementary material](#)



Published online: 18 Sep 2026.



[Submit your article to this journal](#)



Article views: 117



[View related articles](#)



[View Crossmark data](#)

Spatial prediction of groundwater nitrate through machine learning and stakeholder co-creation: the Chalk aquifer case in East Anglia, UK

V. Gómez-Escalonilla^a, H. Baron ^b, T. Read^c, J. Watson^c, M. Rodríguez Del Rosario^a, G. Darch^c, V. Keller^b, N. Rickards^b and P. Martínez-Santos^a

^aDepartamento de Geodinámica, Estratigrafía y Paleontología, Universidad Complutense de Madrid, Madrid, Spain; ^bWater Resources & Drought, UKCEH, Wallingford, United Kingdom of Great Britain – England, Scotland, Wales; ^cAnglian Water Ltd., Huntingdon, UK

ABSTRACT

Nitrate contamination in groundwater threatens drinking water supplies worldwide. Data-driven methods can improve prediction and spatial mapping by transforming point observations into continuous regional information, supporting a better understanding of contamination patterns and mitigation strategies. This study presents the co-creation with stakeholders of a continuous nitrate contamination map for the Chalk aquifers of East Anglia, England. Among the classification models, Gradient Boosting and Random Forest performed best, each achieving an accuracy of 0.89. For regression, the Extra Trees model showed the highest performance ($R^2 = 0.64$; MAE = 11.0 mg/L). Both regression- and classification-based maps captured nitrate distribution effectively, identifying the unconfined Chalk outcrops in the west and parts of the central and northern valleys as the areas of greatest concern, while confined Chalk aquifers showed lower contamination potential. Both approaches also highlighted unsampled locations outside official nitrate vulnerable zones with characteristics similar to known contaminated areas. Shapley analyses identified geological variables as the strongest predictors of groundwater nitrate. Although both mapping approaches produced similar spatial patterns, they provide complementary information that can support stakeholder decision-making, with improved model interpretability enhancing their practical uptake.

ARTICLE HISTORY

Received 9 January 2026
Accepted 15 July 2026

ASSOCIATE EDITOR

Michael Nones

KEYWORDS

Ensemble algorithms; nitrate contamination; regression; classification; groundwater

1. Introduction

Nitrate contamination in natural waters is one of the most pressing environmental challenges of the twenty-first century (Abascal *et al.* 2022). Excess nitrate threatens drinking water and disrupts ecosystem dynamics, leading to eutrophication, biodiversity loss, and long-term degradation of aquatic habitats (Wang *et al.* 2023). Because nitrate is highly soluble and mobile, its presence in the environment is governed by a combination of human practices, hydrological pathways, and geogenic and climatic factors (Rosenstock *et al.* 2014, Mahlknecht *et al.* 2023). Capturing complex interactions among these is crucial to devise effective management strategies for aquifer systems (Spalding and Exner 1993, Gómez-Escalonilla *et al.* 2024). This is, however, far from straightforward, as the processes and lag times that constrain nitrate transport through the unsaturated zone and within aquifer systems are difficult to assess (Wang *et al.* 2013). Furthermore, direct measurement of nitrate at high spatial resolution is often unfeasible due to high fieldwork and laboratory costs, the sparsity of

monitoring wells, and the technical difficulties of long-term sampling. Additional sources of uncertainty in nitrate monitoring include hydrogeological considerations, sampling methods and strategies, and water point typology and construction (García-Aróstegui *et al.* 2024).

The above motivates an increasing reliance on computational models to interpolate, understand, and ultimately predict nitrate concentrations in time and space. Process-based models have long since been used to represent nitrate transport and concentration in soils and aquifers (Jiang and Somers 2009, Eryigit and Engel 2022). While grounded in physical and chemical principles, these approaches rely on strong assumptions about linearity and require extensive parameterisation that can rarely be satisfied with the available data. In this context, the rise of artificial intelligence (AI), and more specifically the machine learning (ML) offers a versatile alternative whose full potential is yet to be explored. Because ML models are adept at establishing non-linear relations between a given target – nitrate concentration – and a set of suitable predictors – land use, topography, soil attributes,

CONTACT P Martínez-Santos  pemartin@ucm.es  Departamento de Geodinámica, Estratigrafía y Paleontología, Universidad Complutense de Madrid, Madrid 28040, Spain

 Supplemental data for this article can be accessed online at <https://doi.org/10.1080/15715124.2026.2706022>.

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

rainfall, concentration of other chemical species, and so on –, ML algorithms can be used to predict nitrate concentration without explicitly relying on physical equations (Bhattarai *et al.* 2021, Chahid *et al.* 2025).

In the recent past, the data-driven paradigm has shown significant promise in many groundwater-related applications, where complex, nonlinear, and spatially heterogeneous systems are the norm (Díaz-Alcaide and Martínez-Santos 2019, Elsayed *et al.* 2025, Martínez-Santos *et al.* 2021). These applications span from mapping groundwater potential (Asadi Nalivan *et al.* 2023) and groundwater-dependent ecosystems (Duran-Llacer *et al.* 2025) to the generation of predictive groundwater-quality maps for specific chemical parameters (Podgorski and Berg 2020, Ransom *et al.* 2022). The outcomes of existing studies suggest that ML has the potential to bridge the gap between sparse field observations and the demand for spatially continuous nitrate maps. In this context, random forest and other tree-based ensemble algorithms have been widely used to estimate nitrate concentrations in groundwater, due to their ability to handle nonlinear relationships, complex interactions among predictors, and heterogeneous datasets, while being relatively robust to noise and overfitting. These models have demonstrated robust performance across heterogeneous landscapes in a variety of geographical contexts (Knoll *et al.* 2019, Ransom *et al.* 2022, Gómez-Escalonilla *et al.* 2024). Neural networks and deep learning approaches also show potential in terms of capturing non-linearities and high-dimensional interactions that are difficult to model with more conventional methods (Karimanzira *et al.* 2023). In addition, the integration of spatial information through geostatistical or convolutional methods has helped ML models account for spatial autocorrelation and neighbourhood effects that strongly influence nitrate distributions (Mehrabi *et al.* 2025). For this study, tree-based ensemble models in both classification and regression, specifically Random Forest (Breiman 2001), AdaBoost (Freund and Schapire 1997), Gradient Boosting (Friedman 2001), and ExtraTrees (Geurts *et al.* 2006), were selected as a robust and well-established approach for the spatial prediction of groundwater nitrate concentrations.

Supervised learning applications distinguish between classification and regression (Pedregosa *et al.* 2011). Regression models are trained to predict a target as a continuous variable, providing fine-grained estimates for concentrations at unsampled locations. In contrast, classification outcomes are categorical. This means classification may be better suited to datasets that include non-numerical predictors such as geological formations and land-use types. Furthermore, policymakers and water managers often operate within threshold-based frameworks, in which the key

question is not the precise nitrate concentration, but whether this concentration exceeds regulatory limits (Singh *et al.* 2022). While both regression and classification are becoming increasingly common in the groundwater contamination literature (Lee *et al.* 2023, Yang *et al.* 2023), the existing studies seldom compare the relative performance of both frameworks on the same dataset. This gap is striking given the rich ecosystem of ML-based methods that have become available to researchers in the last few years. Understanding the implications requires carefully designed experiments in which both approaches are trained, validated, and compared on the same spatial data. Examples such as the one we present in the following pages could shed light on important trade-offs in contexts where the choice between classification and regression might not be obvious at first glance. Such instances include cases with skewed data distributions, when thresholds are of primary concern, or when imbalanced sampling leads to challenges in modelling extremes.

Machine learning models in hydrology traditionally suffered from the lack of interpretability. This hampered attempts to draw scientific insights and limited potential uptake from stakeholders. Since relatively recent times, however, algorithm suites incorporate routines that make the results more transparent to users. Notable examples include local interpretable model-agnostic explanations, which approximate local decision boundaries with interpretable models (Ribeiro *et al.* 2016), and Shapley additive explanation methods (SHAP), which provide a unified, game-theoretic framework for attributing predictions to input features with strong theoretical guarantees (Lundberg and Lee 2017, Lundberg *et al.* 2020). While these types of analyses remain rare in the groundwater literature, they are gradually gaining recognition as tools to enhance the explainability of water-related ML applications (Makumbura *et al.* 2024, Yang *et al.* 2024).

This paper offers a stakeholder-based application of machine learning tools to map groundwater nitrate across East Anglia's Chalk aquifers. Models were developed in close collaboration with the staff of Anglian Water, a water company that supplies drinking water to all or parts of Bedfordshire, Buckinghamshire, Cambridgeshire, Essex, Lincolnshire, Norfolk, Northamptonshire, and Suffolk. Partnerships that combine private-sector expertise with proprietary datasets are still uncommon in the academic literature, where ML-based groundwater studies more often depend on publicly available data and simplified research environments. In this context, we carry out a systematic comparison between regression and classification techniques for predicting nitrate concentrations in space. Our goal is two-fold. Firstly, we evaluate how classification and regression models can be applied to the same dataset of nitrate measurements

and spatial covariates, assessing their relative strengths, weaknesses, and implications in a specific case study. The dual approach acknowledges that scientific research and policy decision-making often demand different kinds of outputs, and highlights how ML can be adapted to serve both. Secondly, by applying these methods to predict nitrate across East Anglia's Chalk aquifers, we aim to further our understanding of the processes that govern nitrate contamination in the study area. To do so, we leverage an unusually comprehensive dataset that combines existing hydrogeological knowledge with state-of-the-art interpretability methods, including feature importance, partial dependence, and SHAP analyses, thereby advancing both methodological and applied insights in a domain where such integrated approaches are still limited.

2. Materials and methods

2.1. Geographical and geological context

The study area extends across 14,000 km², covering parts of Lincolnshire, Norfolk, Suffolk, Essex, and Cambridgeshire (Figure 1). This is a key arable crop-growing region of England and is home to many important wetland ecosystems and a variety of ecological river habitats. The principal sources of water supply in the region are aquifers, rivers and pump filled reservoirs.

More specifically, we focus on parts of the Chalk aquifer, a Principal Aquifer of southern and eastern England. A principal aquifer is defined as one that is able to store and transmit regionally important quantities of water, supporting public water supply, industry, agriculture and the environment. The Chalk is present in the south east of Yorkshire, southwards across the Humber and into Lincolnshire. It is also found east and south of the Wash in a swathe across central southern England from north Norfolk, through the Thames Basin and along the Kent coast, down to the south coast, the Isle of Wight and into Dorset towards Portland Bill (BGS 2025). The Chalk aquifer is characterised by a microporous white limestone with low matrix permeability and a well-developed network of fractures. The highest borehole yields are in the order of 150 l/sec. These occur where the density of fractures is greatest and/or where the fractures are enlarged, typically in the top 80–100 m. The aquifer is most vulnerable to surface pollutants where the Chalk outcrops at the surface, in areas where there is little or no lower-permeability geological cover.

The study area is divided in two sub-regions, separated by The Wash estuary. The Chalk aquifer is more commonly unconfined along the northwestern edge in the southern part of the southern sub-region, where the lower permeability superficial cover increases in thickness towards the southeast. In the northern

sub-region, around Lincolnshire, this is slightly different, as the aquifer outcrops along the southwestern edge with increasing lower permeability superficial geology towards the northeast. The bedrock and superficial geology across the study area is illustrated in Figure 2.

The predominant land use overlying the Chalk aquifer is agriculture (Figure 1), with arable agriculture forming a large proportion of the land use. The largest areas of non-agriculture of note include Thetford Forest and The Brecklands. Agricultural land includes the application of organic and inorganic fertiliser. The residual nutrients post-harvest can leach quickly from well-draining soils during the first autumn flush. This effect, coupled with the East of England receiving lower volumes of rainfall, leads to a higher leached nitrate concentration when compared to the same mass of nitrate loss in wetter parts of the country.

2.2. Supervised machine learning approach

Supervised learning is a specific subset of machine learning in which a model is trained on a dataset to predict the outcome of a target variable. The dataset consists of a series of entries in which both the value of the target variable and of the predictor variables are known. Algorithms then identify patterns that associate predictors with the target variable, to make predictions in cases where the target outcome is unknown. There are two major supervised learning approaches, namely, classification and regression. The difference is that classification works with a categorical target, whereas regression attempts to predict the value of a continuous variable.

Initial testing for collinearity was conducted on all explanatory variables. Collinearity, defined as the existence of a strong correlation between two explanatory variables, can be a key consideration within this framework. Indeed, pairwise collinearity may result in overemphasising the importance of a given predictor, and may also incorporate unwanted noise to the results. In this study, multicollinearity was evaluated using the Variance Inflation Factor (VIF). This metric quantifies how much the variance of a regression coefficient is inflated due to correlations with other predictors in the model. Low VIF values indicate little multicollinearity, whereas higher values suggest increasing redundancy among variables. Following commonly adopted thresholds in environmental modelling studies, variables with VIF values greater than 10 were considered to exhibit problematic levels of multicollinearity and were evaluated for potential removal from the predictor set (Dormann *et al.* 2013).

Figure 3 shows a schematic overview of the workflow. Steps include data preparation, model training and evaluation, model interpretation, and

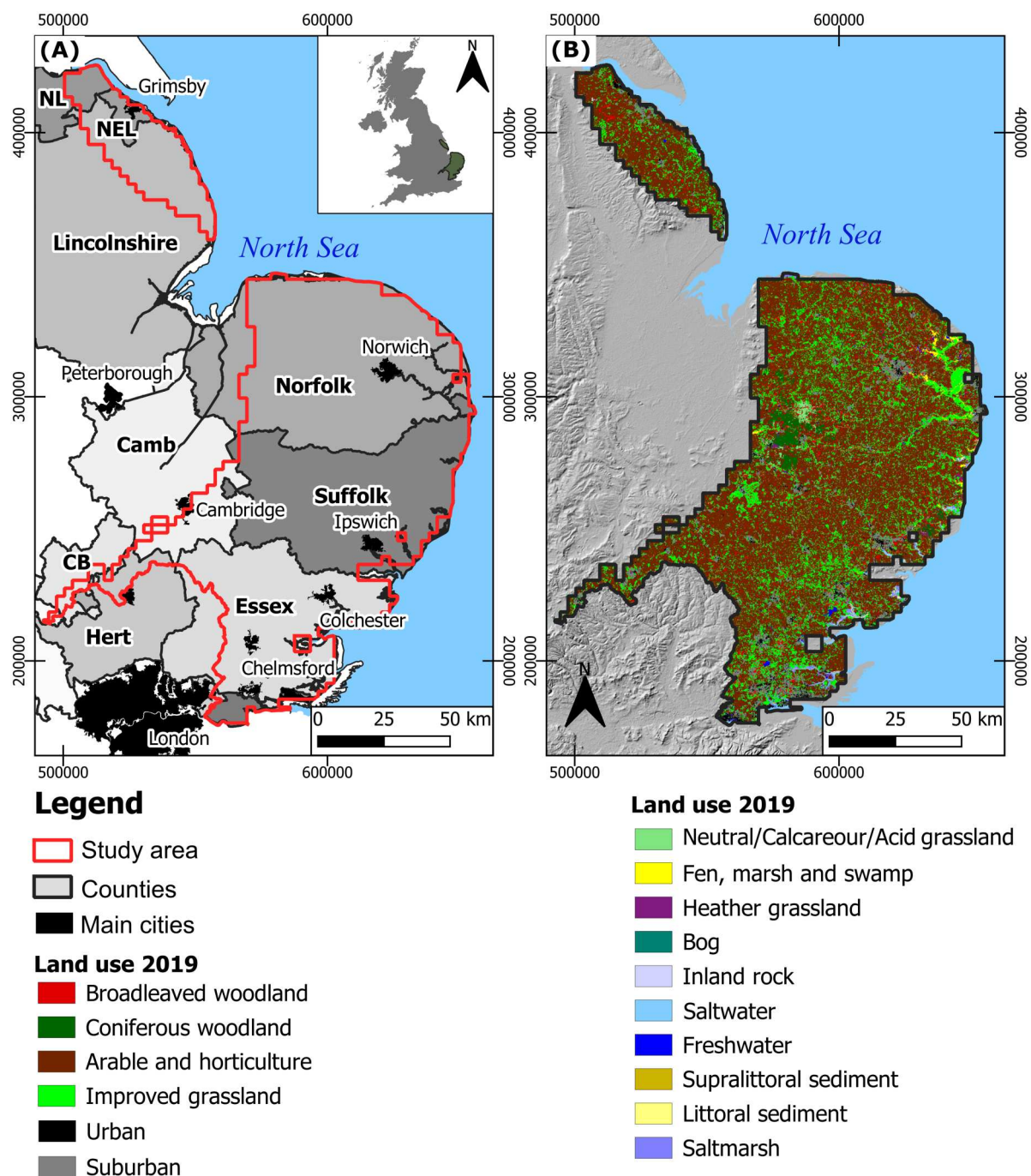
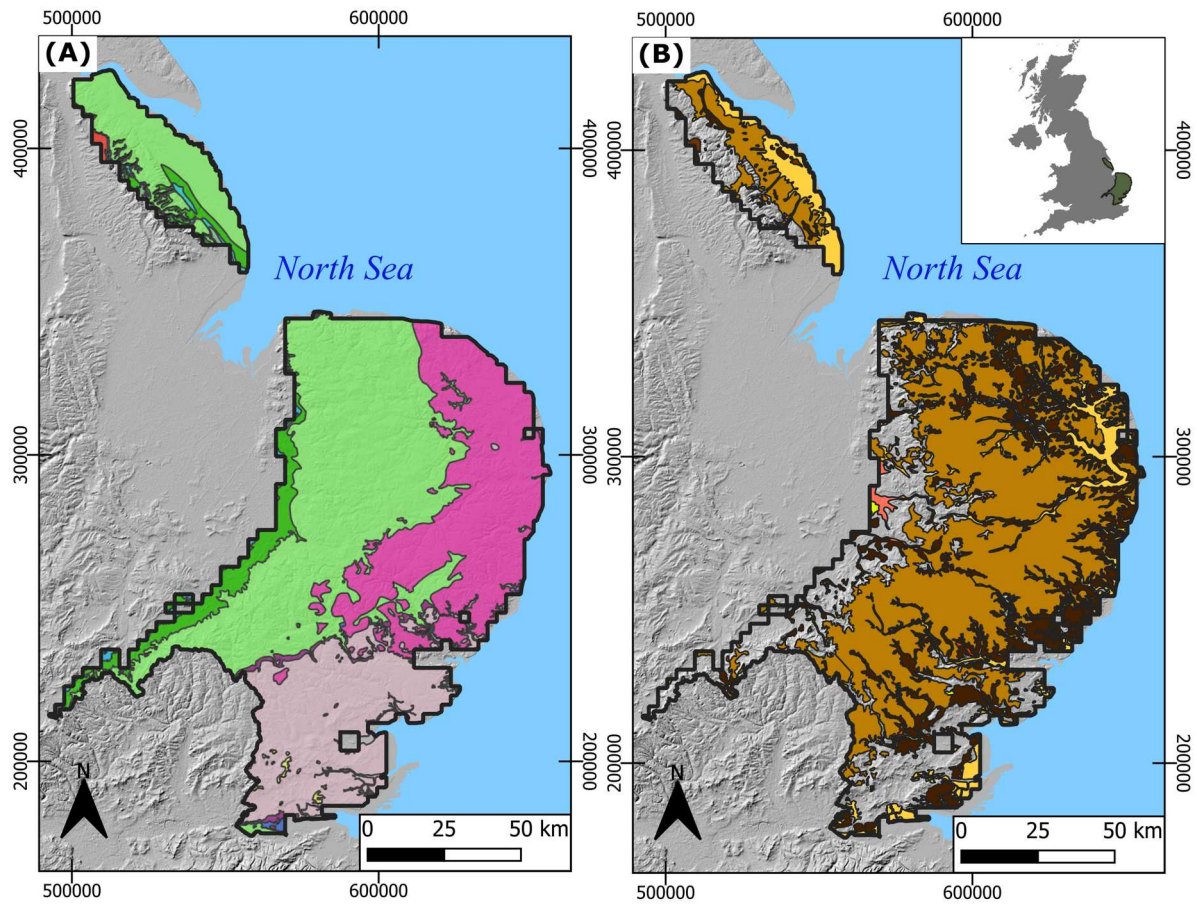


Figure 1. Location and land use setting of the study area showing (A) Administrative regions (counties) and main cities and (B) land use distribution in 2019. Counties abbreviations: NL: North Lincolnshire; NLE: North East Lincolnshire; Camb: Cambridgeshire; CB: Central Bedfordshire; Hert: Hertfordshire. Land cover based on Land Cover Map 2019 © UKCEH 2020. Contains Ordnance Survey data © Crown Copyright 2007, Licence number 100017572 (Morton *et al.* 2020).

finally, the generation of spatial predictive maps. Model training and evaluation were conducted using a repeated nested cross-validation framework, an approach designed to guarantee robust and unbiased performance estimates with less dependence on the random train/test split (Cawley and Talbot 2010, Krstajić *et al.* 2014). Specifically, a repeated stratified k-fold scheme (5 folds with 10 repetitions) was implemented as the outer loop, generating 50 independent train–test splits. Within each outer iteration, the training subset was used in an inner stratified k-fold cross-validation (3 folds) to optimise

model hyperparameters using a Randomized-SearchCV procedure. The hyperparameter search space included `min_samples_split`, `min_samples_leaf`, `max_depth`, `n_estimators`, `learning_rate`, and `subsample`, depending on the algorithm. In this procedure, 20 iterations were used to obtain the optimal values. This value was obtained through a prior sensitivity analysis that showed that an increase in the number of iterations was not associated with a positive impact on model performance, while significantly increasing computational cost. The best hyperparameter configuration identified in the



Legend

STUDY AREA

Bedrock geology

- BRACKLESHAM GROUP AND BARTON GROUP
- GAULT FORMATION AND UPPER GREENSAND FORMATION
- GREY CHALK SUBGROUP - CHALK
- WHITE CHALK SUBGROUP - CHALK
- LAMBETH GROUP
- LOWER GREENSAND GROUP
- NEOGENE ROCKS
- NEOGENE TO QUATERNARY ROCKS
- THAMES GROUP
- THANET SAND FORMATION
- WEALDEN GROUP
- WEST WALTON FORMATION, AMPTHILL CLAY FORMATION AND KIMMERIDGE CLAY FORMATION

Superficial geology

- CLAY
- CLAY, SILT AND SAND
- DIAMICTON
- PEAT
- SAND
- SAND AND GRAVEL
- SILT

Figure 2. Location and geological setting of the study area showing (A) bedrock geology and (B) superficial geology. Geological maps of the bedrock and surface horizons obtained from data provided by the British Geological Survey © UKRI [2026] (BGS 2016a).

inner loop was then used to retrain the model on the full outer training set and subsequently evaluated on the corresponding outer test set. This approach prevents information leakage between training and evaluation stages and provides reliable estimates of model generalisation performance. Definitions and optimised values for these hyperparameters for the

models are provided in Table S1. Additionally, recursive feature elimination with cross-validation (Pedregosa *et al.* 2011) was explored as a feature selection strategy; however, sensitivity analysis showed that it significantly increased computational time without yielding improvements in predictive performance, and was therefore not included in the final workflow.

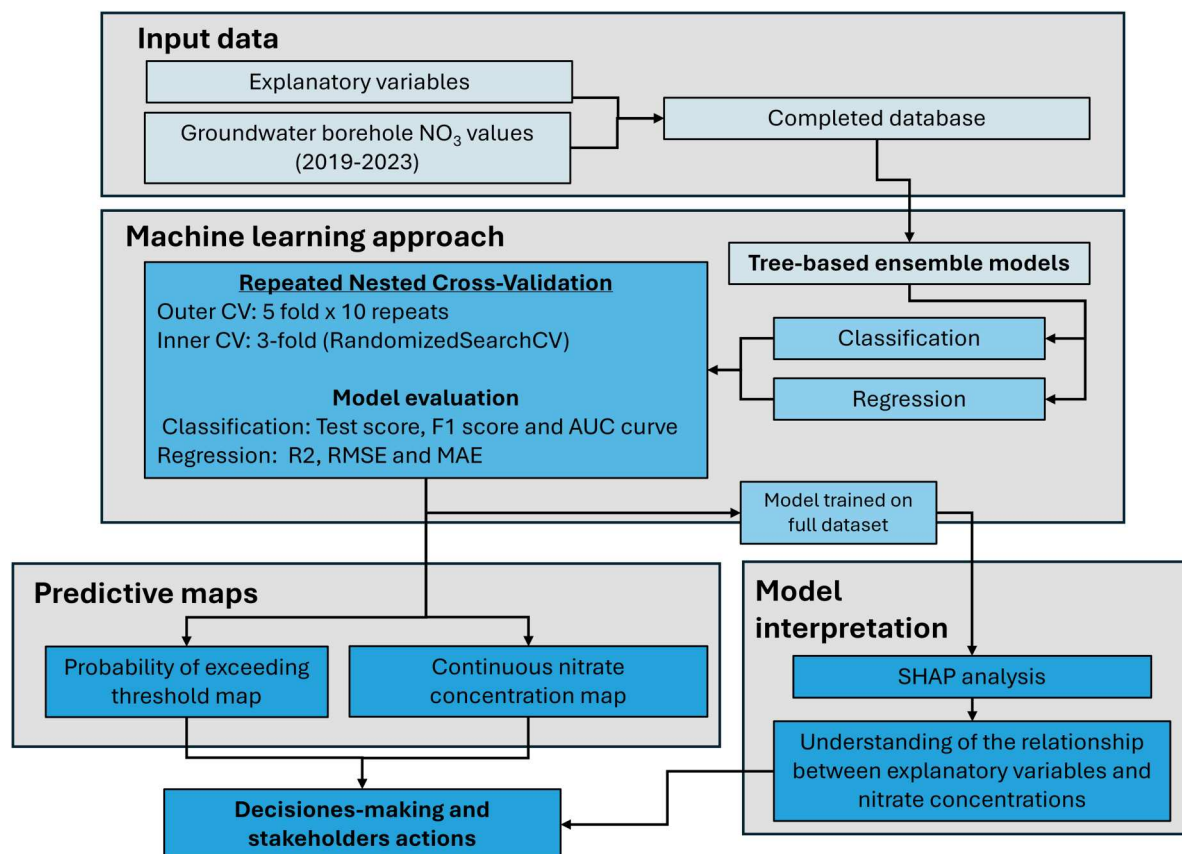


Figure 3. Schematic overview of the modelling workflow.

Traditionally, machine learning approaches have been subject to criticism due to their perceived lack of transparency and interpretability. For this reason, there has been an increase in the number of methodologies aimed at improving this in recent years. We use the SHAP strategy (Lundberg *et al.* 2020) with the intent of corroborating machine learning results with the conceptual hydrogeological model of the study area, and, ultimately, of increasing the value of the outcomes in decision-making. SHAP facilitates a deeper and more comprehensive understanding of the model's predictions by calculating feature importance scores. In contrast to conventional approaches that evaluate features independently, SHAP is based on Shapley values from game theory. This approach evaluates the contribution of each feature by considering all possible combinations, which provides an integrated perspective as to how features interact and collectively influence model decisions (Bilgilioglu *et al.* 2025). Once the models were evaluated using the nested cross-validation framework, a final model was trained with the sole objective of interpreting the role played by the covariates. For this phase, each algorithm was fitted using the entire dataset, since the validation process had already been completed and unbiased performance estimates had been obtained. This approach ensures that the SHAP analysis is based on the maximum amount of available information, allowing for a more stable and

representative assessment of the features' contributions without compromising the integrity of the model evaluation.

Finally, classification and regression algorithms were used to generate spatial predictions of nitrate pollution across the study area. The classification model produced maps representing the probability of exceeding the threshold defined in the target variable, whereas the regression model generated continuous nitrate concentration estimates based on the selected predictors. For the classification approach, the final predictive map corresponds to the average probability obtained from the ensemble of the 50 models generated through the repeated nested cross-validation procedure, considering all four tree-based algorithms. This aggregation strategy provides more robust and reliable spatial predictions while also allowing the quantification of model uncertainty. In the case of regression, the final map was derived from the best-performing model.

2.2.1. Classification

Ensemble tree-based classification algorithms have proven particularly effective in predictive mapping tasks in the past (Heung *et al.* 2014, Rodriguez-Galiano *et al.* 2014, Mosavi *et al.* 2021, Zounemat-Kermani *et al.* 2021). These models are all based on simple decision trees as weak learners, but use different strategies to combine the prediction of the models to

improve the accuracy and robustness of the results. Random Forest Classifier (RFC), AdaBoost Classifier (ABC), Gradient Boosting Classifier (GBC), and ExtraTrees Classifier (ETC) were selected in this case. All of them originate from the SciKit-Learn toolbox (Pedregosa *et al.* 2011) and were integrated within the MLMapper 2.0 software (Gómez-Escalonilla *et al.* 2022a, 2022b).

The Random Forest Classifier (RFC) is an ensemble learning method that combines multiple decision trees built from bootstrap samples of the training data (Breiman 1996, 2001). By introducing randomness in data sampling and feature selection, RFC effectively reduces variance and overfitting. This results in improved predictive accuracy and model stability compared to individual decision trees (Pedregosa *et al.* 2011). The ExtraTrees Classifier (ETC) operates in a similar manner, but trains all trees on the complete dataset without bootstrap sampling and introduces additional randomness by selecting attribute splits and thresholds entirely at random. This approach slightly increases bias but effectively reduces variance and substantially accelerates model training (Geurts *et al.* 2006, Géron 2017).

In contrast, boosting methods such as the AdaBoost Classifier (ABC) and the Gradient Boosting Classifier (GBC) build models sequentially rather than in parallel. The ABC algorithm converts weak learners into strong classifiers by iteratively adjusting sample weights to correct misclassifications, combining the resulting learners through weighted voting based on their predictive performance (Freund and Schapire 1997, Pedregosa *et al.* 2011, Géron 2017). The GBC, on the other hand, sequentially fits each new tree to the residual errors of the preceding model, progressively reducing prediction errors and enhancing performance through iterative refinement (Friedman 2001, Géron 2017).

2.2.2. Regression

The regression models operate in a similar manner to the classification models described above, but with the models being trained on a continuous variable rather than a categorical one, therefore yielding a numerical value as a prediction. The models investigated include Random Forest Regressor (RFR); AdaBoost Regressor (ABR); Gradient Boosting Regressor (GBR); and the ExtraTrees Regressor (ETR).

In the literature, classification algorithms are generally applied to predictive mapping problems while regression algorithms are often used for temporal predictions. However, there are several examples where regression algorithms have been successfully applied for spatial predictions, including RFR for predicting nitrate in groundwater in China (He *et al.* 2022) and fluoride concentration in the western United States (Rosecrans *et al.* 2022). In Spain, multiple ensemble

tree regressors were used to predict groundwater level, both spatially and temporally (Martínez-Santos *et al.* 2025).

2.2.3. Selection of performance metrics

The type of target variable predicted by classifiers and regressors is different. This means that different metrics must be used to evaluate the performance of machine learning models. The main metrics considered for the classifiers used in this work were the test score, the F1 score for both classes, and the AUC curve, while the metrics considered for evaluating the regression-based models were the coefficient of determination (R^2), the Root Mean Square Error (RMSE), and the Mean Absolute Error (MAE). For the regression-based models, the metrics were obtained for both the training and test data of each outer loop. For both approaches, the mean and standard deviation of the metrics obtained from the 50 models were considered.

Test score is the simplest from a conceptual standpoint. It is calculated as the number of correct attempts over the total number of samples in the test data set. This metric gives a first approximation about the predictive power of the models. However, it can also yield biased results on unbalanced datasets. For this reason, the F1 metric was used to analyse the quality of the predictions in each of the concentration classes separately. This metric is calculated as the harmonic mean of precision and recall, where precision is the ratio of true positives over the sum of true and false positives and recall is the ratio of true positives over the sum of true positives and false negatives. Finally, the AUC curve, a probabilistic metric, illustrates how the diagnostic ability of a binary classifier system varies with its discrimination threshold. That is, it measures how well a given algorithm can distinguish between the two classes in a binary problem. AUC values range between 0 and 1.

The R^2 metric is commonly used to evaluate model performance and measures the proportion of variance in the target variable which can be explained by the model. R^2 varies between 0 and 1, with 1 indicating a perfect model. The RMSE is a measure of the differences between the predicted and observed values, and MAE is the average magnitude of errors between predicted and observed values. Both RMSE and MAE vary between 0 and infinity, with 0 indicating a perfect fit.

2.3. Data sources

2.3.1. Borehole database description

Groundwater nitrate concentration data was obtained from Anglian Water's supply boreholes. The database contains 483 boreholes drilled between 1898 and 2013, across seven different principal aquifers. Key asset

information included the year of borehole installation, diameter, total depth, open/slotted interval depth range, aquifer, and the depth to the top of the aquifer. The nitrate concentration time series available for this study span a 10-year period, between 2013 and 2024.

Boreholes where the intake is within the Chalk aquifer were extracted from this database and taken forward. These 283 boreholes ranged from 11 to 160 m depth. This range reflects the feather edge of the Chalk in North West Norfolk, and the deeper chalk found in the central and eastern parts of the southern sub-region, and in parts of Lincolnshire. Similarly, the top of the Chalk aquifer varied from 0 to 70 m depth. This reflects the different hydrogeological contexts found from the eastern part, where the aquifer is unconfined, to areas where it is overlain by sands-gravels-clays, and deeper confined conditions in the south east of the region, beneath the Thames Formation.

2.3.2. Stakeholder-based definition of the target variable

Predictive models were co-developed with personnel from Anglian Water. This was a collaborative process in which researchers and practitioners participated actively in designing, developing, and evaluating the experiment. Our basic intent was to provide a continuous map of nitrate contamination across the two sub-regions that make up the study area. More specifically, we attempted to evaluate the capabilities of ML algorithms to depict known hydrogeological conditions and to extrapolate the outcomes to unsampled locations. Furthermore, we examined the extent to which regional results could be downscaled to explain differences in nitrate concentration between boreholes located in close proximity to each other.

For this we used supervised classification and regression approaches. With regard to classification, establishing the threshold to distinguish between contaminated and uncontaminated is a fundamental first step. Several scenarios were evaluated during the data preparation phase. This allowed us to explore the existing time series, working with 25, 37.5, and 50 mg/L cutoff values as per the UK and World Health Organization's drinking standards, the EU's Nitrate Directive, and reference values from different EU Member States. The 50 mg/L option was subsequently determined to be the most appropriate for local stakeholders, as it reflects the legal threshold for nitrate contamination in drinking water supplies in the UK. This value also provides a well-balanced split of contaminated and uncontaminated data points.

Given the availability of long time series and the objective of enhancing the representativeness of the results, the analysis was restricted to the last five years (2020–2024). Only boreholes that comprised a minimum of ten samples for this timeframe were

considered. Out of these, only those with a nitrate concentration exceeding 50 mg/L in more than 10% of the samples were considered contaminated. Any boreholes with a content in excess of 50 mg/L in less than 10% of the samples, or that never reached this value during the 2020–2024 interval, were labelled uncontaminated.

For the regression approach, a continuous variable is required for the target variable. For this work it was decided that the median value of nitrate concentration was the most representative metric. As in the classification approach, boreholes with fewer than 10 samples were not included. Figure 4 shows the spatial distribution of the data for both approaches.

2.3.3. Spatial predictors

Spatial predictors are spatially-distributed variables that can be expected to explain groundwater contamination to a certain extent. These have been extensively discussed in the literature (Knoll *et al.* 2019), and include geological and hydrogeological properties, topographical factors, soil related variables, climatic information, and potential nitrogen sources.

Land use is a major predictor of groundwater nitrate, as agricultural activities, landfills and water discharge points are frequent sources of contamination (Lerner and Harris 2009). Nitrate subsequently enters aquifers through irrigation and/or rainfall, which in turn influence groundwater flow patterns, nitrate leaching, nitrate dilution, and subsurface transport (Sun *et al.* 2024, Zhu *et al.* 2024). Additionally, infiltration is controlled by soil composition and texture, as well as by topographic factors (Gaines and Gaines 1994, Díaz-Alcaide and Martínez-Santos 2019). Finally, from a hydrogeological perspective, nitrate concentration is dependent on the hydrodynamic properties of the rock, on the presence of confining or semi-confining layers, and on the existence of oxidising or reducing conditions. These factors may speed up or delay transport, and can also alter nitrification and denitrification processes (Rivett *et al.* 2007).

In this case, predictors were selected based on the literature (Knoll *et al.* 2019, Spijker *et al.* 2021, Kari-manzira *et al.* 2023, Mahboobi *et al.* 2023), as well as on the experience of the stakeholders. Table 1 presents a summary of all predictors.

3. Results

3.1. Collinearity analysis

Multicollinearity among the predictor variables was assessed using the Variance Inflation Factor (VIF), and the results are presented in Table 2. The analysis revealed high multicollinearity ($VIF > 10$) among several variables. In particular, strong collinearity was

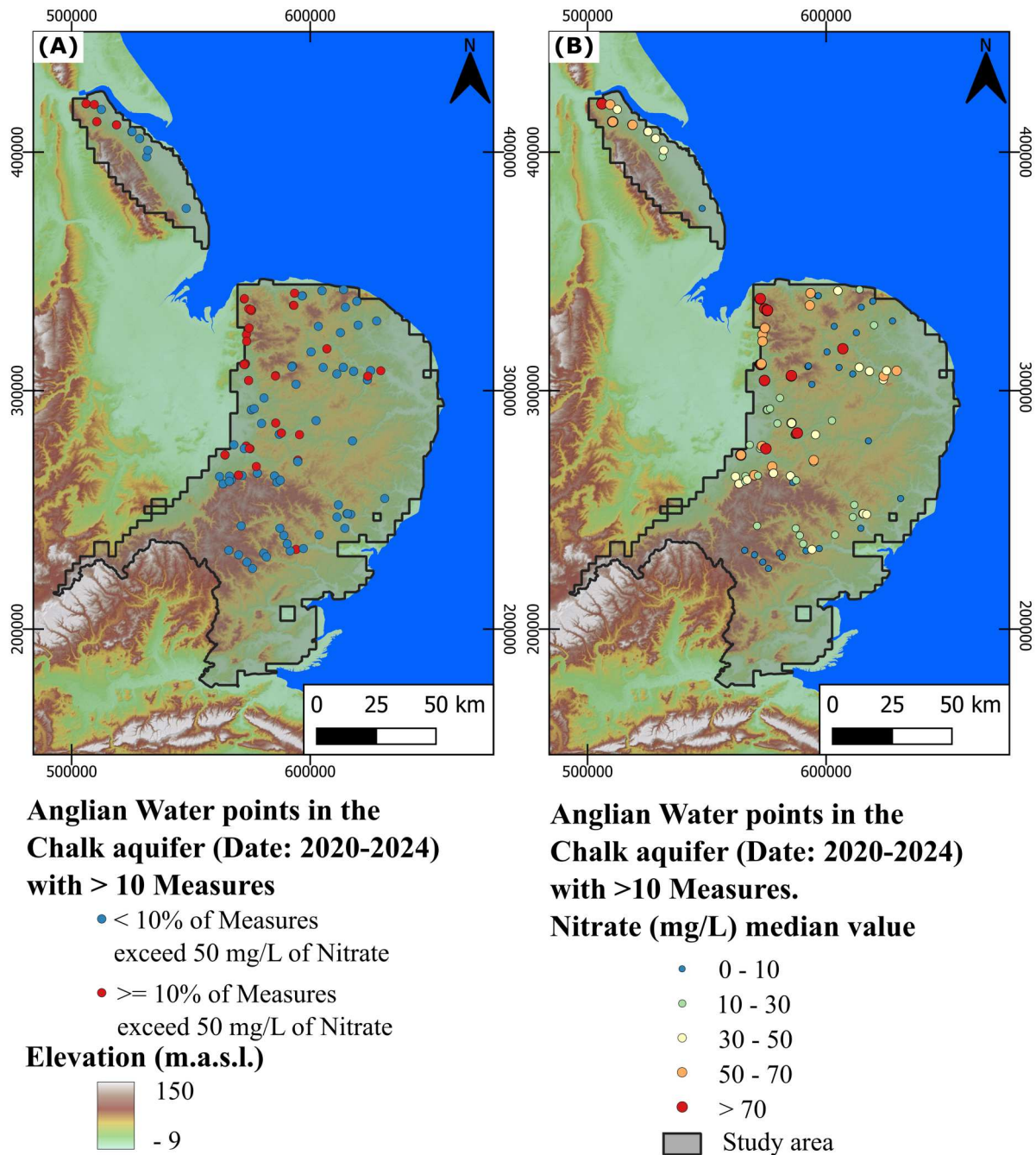


Figure 4. Spatial distribution of the data used in the two supervised learning approaches: (A) Locations of boreholes used in the classification approach. (B) Locations of boreholes coloured by median nitrate concentrations (mg/L) during the study period. These concentrations served as input data for the regression approach. Data obtained from Anglian Water's supply boreholes.

identified between Elevation, GWL_Elevation, and UZ_Thickness. This is expected because UZ_Thickness is derived from the difference between Elevation and GWL_Elevation, which inherently produces high VIF values. A strong relationship was also observed between Thickness_Q_materials and Buried_valleys. Although these variables originate from different datasets, they represent closely related geological-geomorphological processes, as buried valleys are the main zones where Quaternary sediments tend to accumulate.

To reduce redundancy and improve the robustness of the modelling framework, the variables UZ_Thickness and Buried_valleys were removed from the

predictor set. These variables were considered less relevant for the modelling objectives, had lower spatial resolution, or represented derivative information from other predictors already included in the dataset. After their removal, VIF values were recalculated, showing a substantial decrease and indicating that multicollinearity had been largely mitigated.

Some degree of correlation remained between Elevation and GWL_Elevation. However, this relationship reflects a well-known hydrogeological control, as groundwater levels are strongly influenced by topography. Therefore, both variables were retained in the final predictor set in order to preserve the physical representativeness of the hydrogeological system.

Table 1. Summary of the data used as spatial predictor variables (BGS: British Geological Survey, EA: Environment Agency).

Type	Spatial predictors (variable name)	Description	Resolution	Source – Reference
Topographical factors	Topographic wetness index (TWI)	Topographic wetness index (TWI) elaborated from DEM	30 m	Derived data
	Slope (Slope)	Terrain slope elaborated from DEM	30 m	Derived data
	Landforms (Geomorphon)	Landforms elaborated from DEM using r.geomorphon	30 m	Derived data
	Elevation (Elevation)	Digital elevation model from SRTM (Shuttle Radar Topography Mission)	30 m	SRTM ^a
Hydrogeological factors	Distance to channels (Rivers_distance)	Distance to main rivers (statutory watercourses) designated by the Environment Agency as main rivers.	–	EA (2023)
	Distance to lakes (Lakes_dist)	Distance to lakes lake polygons from the UK Lakes Portal	–	Taylor (2021) ^b
	Superficial geology (Superficial_geology)	Deposits laid down during the Quaternary Period	1:50.000	BGS (2016a) ^c
	Bedrock geology (Bedrock_geology)	Rocks and deposits laid down prior to 2.588 million years ago- but including the Crag Group	1:50.000	BGS (2016a) ^c
	Distance to Fractures (Distance_Faults)	Distance to the fractures	1:250.000	Kearsey et al. (2019) ^d
	Buried valley thickness (Buried_valleys)	Ancient river or subglacial drainage networks, now abandoned and partly or completely buried by more recent sediment. As such, buried valleys often exhibit little or no surface expression within the modern landscape.	1 km	BGS (2016b) ^e
Soil-related factors	Thickness of superficial deposits (Thickness_Q_materials)	The superficial thickness model 1 km hex grid shows the variation of the thickness of superficial (Quaternary age) deposits across Great Britain.	3 km	BGS ^f
	Chalk Aquifer base elevation (Chalk_Base)	Chalk Aquifer base elevation obtained by means of a geological modelling tool (GS3D) developed by the BGS	1:50.000	BGS (2021a) ^g
	Superficial materials permeability (Superf_permeability)	Permeability is based on geological considerations for superficial materials.	1:50.000	BGS (2021a) ^g
	Bedrock materials permeability (Bedrock_permeability)	Permeability is based on geological considerations for bedrock materials.	1:50.000	BGS ^h
	Unsaturated Zone Thickness – Groundwater depth (UZ_Thickness)	The depth to groundwater dataset is a gridded representation of the depth, in metres, from the ground surface to the phreatic water table.	50 m	BGS ^h
	Groundwater Level Elevation (GWL_Elevation)	Derived from Elevation and Depth to groundwater	50 m	BGS ^h
	Soil parent material (Soil_Parent)	Soil parent material dataset informs about the likely soil and sub-soil characteristics. The data provides information about ground conditions relating to agricultural soils.	1:50000	BGS (2021b) ⁱ
	Sand content (SAND_MEAN)	Proportion of sand particles (>0.05/0.063 mm) in the fine earth fraction (g/kg) in the upper 2 m of soil.	250 m	SoilGrids – Poggio et al. (2021)
	Clay content (CLAY_MEAN)	Proportion of clay particles (<0.002 mm) in the fine earth fraction (g/kg) in the upper 2 m of soil.	250 m	SoilGrids – Poggio et al. (2021)
	Coarse fragment content (CFVO_MEAN)	Volumetric fraction of coarse fragments (>2 mm) in the upper 2 m of soil.	250 m	SoilGrids – Poggio et al. (2021)
Climatic factors	Average Rainfall (Rain_aver_2010_2024)	Average annual rainfall (2010–2023) from HadUK-Grid. Gridded climate variables derived from the network of UK land surface observations	1 km	Met Office et al. (2024)
	Average Temperature (Temp_aver_2010_2024)	Average annual temperature (2010–2023) from HadUK-Grid. Gridded climate variables derived from the network of UK land surface observations	1 km	Met Office et al. (2024)

(Continued)

Table 1. Continued.

Type	Spatial predictors (variable name)		Description	Resolution	Source – Reference
Land cover / potential N sources	Land Cover (LC_2019_5class)		Land Cover Map 2019 produced at the UK Centre for Ecology & Hydrology by classifying satellite images from 2019	20 m	Morton <i>et al.</i> (2020) ^j
	Distance to agricultural areas (Distance_agriculture)		Derived from Land Cover Map 2019	30 m	Derived data
	Agriculture type classification (ALC_GRADE_)		Agricultural land classified into five grades. Grade one is best quality and grade five is poorest quality.	1:250,000	Natural England ^k
	Density of Consented Discharge Points (DENSITY_CDP)		Density of Consented Discharge Points Into GW and Land Infiltration	30 m	EA ^l
	Distance to Consented Discharge Points (DISTANCE_CDP)		Distance to the closest Consented Discharge Points. Into GW and Land Infiltration	30 m	EA ^l
	Distance to Landfill sites (Landfill_distance)		Distance to Permitted Waste Sites – Authorised Landfill Site Boundaries	30 m	EA ^m

^a<https://www.usgs.gov/centers/eros/science/usgs-eros-archive-digital-elevation-shuttle-radar-topography-mission-srtm-1>.

^b<https://eip.ceh.ac.uk/apps/lakes/>.

^c<https://www.bgs.ac.uk/datasets/bgs-geology-50k-digmappgb/>.

^d<https://www.bgs.ac.uk/datasets/buried-valleys/>.

^e<https://www.bgs.ac.uk/datasets/superficial-thickness-model-1-km-hex-grid/>.

^f<https://www2.bgs.ac.uk/groundwater/shaleGas/aquifersAndShales/maps/aquifers/home.html>.

^g<https://www.bgs.ac.uk/datasets/permeability/>.

^h<https://www.bgs.ac.uk/datasets/depth-to-groundwater/>.

ⁱ<https://www.bgs.ac.uk/datasets/soil-parent-material-model/>.

^j<https://www.ceh.ac.uk/data/ukceh-land-cover-maps>.

^k<https://naturalengland-defra.opendata.arcgis.com/datasets/Defra:provisional-agricultural-land-classification-alc-england/about>.

^l<https://environment.data.gov.uk/dataset/5fe5ab2e-d465-11e4-8a42-f0def148f590>.

^m<https://www.data.gov.uk/dataset/ad695596-d71d-4cbb-8e32-99108371c0ee/permited-waste-sites-authorised-landfill-site-boundaries1>.

Table 2. Variance Inflation Factor (VIF) values before and after removal of multicollinear variables.

Variable	VIF (Initial Model)	VIF (Final Model)
CLAY_MEAN	4.43	4.19
CFVO_MEAN	2.92	2.92
SAND_MEAN	6.99	6.57
TWI	1.97	1.93
Geomorphon	1.68	1.65
Slope	2	1.75
Elevation	414.65	9.98
Chalk_Base	4.35	4.33
Superf_permeability	2.29	2.16
Bedrock_permeability	2.97	2.82
Superfic_geology	2.33	2.31
Distance_Faults	1.51	1.5
Bedrock_geology	3.82	3.75
Thickness_Q_materials	13.7	3.81
Buried_valleys*	9.85	–
GWL_Elevation	297.86	10.42
UZ_Thickness*	64.75	–
Soil_Parent	1.66	1.62
Lakes_dist	1.66	1.6
LC_2019_5class	1.29	1.27
Distance_agriculture	2.7	2.6
Rain_aver_2010_2024	3.43	3.36
Temp_aver_2010_2024	3.03	2.95
Rivers_distance	2.64	2.64
ALC_GRADE_	4.09	4.06
DENSITY_CDP	2.19	2.08
DISTANCE_CDP	2.32	2.3
Landfill_distance	1.98	1.83

Note: Buried_valleys and UZ_Thickness were the discarded variables and are marked with an asterisk.

3.2. Model performance

Table 3 presents the performance metrics obtained by the four tree-based classification algorithms. All models demonstrated strong predictive performance, as reflected by the high overall test scores and consistent class-wise metrics. In particular, high F1-scores for both the positive (class 1 – ‘contaminated’) and negative (class 0 – ‘uncontaminated’) categories indicate that the classifiers achieve a balanced performance for both contaminated and uncontaminated labels.

Gradient Boosting Classifier (GBC) and Random Forest Classifier (RFC) demonstrates the best overall performance, exhibiting a high generalisation ability (test score = 0.89) and balanced F1-scores for both classes (0.90 for class 0 and 0.87 class 1), along with a high AUC value of 0.95 that confirms its strong capability to discriminate between the two classes. The AdaBoost Classifier (ABC) also achieves robust results, with a test score of 0.87, F1-scores of 0.89

Table 3. Summary of the classification performance metrics (test score, F1-score, and AUC) obtained for the tree-based ensemble models using the repeated nested cross-validation scheme.

Algorithm	Test score	F1 score 0	F1 score 1	AUC
Random Forest Classifier	0.89 ± 0.05	0.90 ± 0.05	0.87 ± 0.06	0.95 ± 0.03
AdaBoost Classifier	0.87 ± 0.06	0.89 ± 0.05	0.85 ± 0.07	0.95 ± 0.04
Gradient Boosting	0.89 ± 0.06	0.90 ± 0.05	0.87 ± 0.07	0.95 ± 0.04
ExtraTrees Classifier	0.86 ± 0.05	0.88 ± 0.04	0.82 ± 0.07	0.95 ± 0.03

and 0.85 for classes 0 and 1, respectively, and an AUC of 0.95. Finally, the Extra Trees Classifier (ETC) presents slightly lower generalisation performance, although it maintains a relatively high score of 0.86. This suggests a greater sensitivity to class imbalance or data variability compared to the other models, which is also reflected in their less balanced F1-scores, indicating comparatively weaker overall performance. For all classifiers, the F1 scores were consistently slightly higher for class 0 than for class 1, indicating that the models performed better at identifying uncontaminated samples. The lower F1 scores for Class 1 could suggest a reduced ability to detect this class, which is likely due to the greater complexity associated with contamination when dealing with more point-specific factors. The standard deviations associated with the various performance metrics were generally low, confirming the robustness of the classifiers evaluated. Random Forest showed the most stable performance, with low variability in the test score (± 0.05), F1 scores (± 0.05 – 0.06), and AUC (± 0.03). Gradient Boosting and AdaBoost showed slightly greater variations, especially for the minority class (F1 score 1, ± 0.07), indicating greater sensitivity to training subsets. However, all algorithms maintained consistent AUC values (± 0.03 – 0.04), demonstrating a stable ability to discriminate between classes across all validation folds.

Table 3 presents the performance of the regression models, evaluated using the R^2 , RMSE and MAE metrics, for the four tree-based ensemble algorithms. Performance is evaluated for both the training and testing phases, where the training results correspond to the models fitted during the outer cross-validation training subsets after hyperparameter optimisation, while the testing results represent the independent outer-loop validation performance. Among the evaluated algorithms, the ExtraTrees Regressor (ETR) achieved the best overall predictive performance, obtaining the highest testing R^2 (0.64 ± 0.11) and the lowest prediction errors (RMSE = 14.54 ± 2.22 mg/L; MAE = 11.05 ± 1.63 mg/L). AdaBoost Regressor (ABR) showed comparable performance ($R^2 = 0.62 \pm 0.12$; RMSE = 14.84 ± 2.52 mg/L; MAE = 11.75 ± 1.73 mg/L), followed by Gradient Boosting Regressor (GBR) and Random Forest Regressor (RFR), which exhibited slightly lower predictive accuracy. Although GBR achieved relatively high training performance, its larger variability across folds indicates reduced stability compared with the other algorithms.

According to the test metrics, the ranking of model performance differed between regression and classification approaches, with Gradient Boosting and Random Forest achieving the best results for classification, whereas AdaBoost and ExtraTrees showed the highest predictive performance for regression.

Table 4. Summary of the regression performance metrics (R^2 , RMSE, and MAE) obtained for the tree-based ensemble models using the repeated nested cross-validation scheme.

Algorithm	R^2		RMSE (mg/L)		MAE (mg/L)	
	Train	Test	Train	Test	Train	Test
RFR	0.82 ± 0.05	0.57 ± 0.11	10.49 ± 1.39	15.86 ± 2.09	8.07 ± 1.20	12.18 ± 1.58
ABR	0.89 ± 0.01	0.62 ± 0.12	8.36 ± 0.41	14.84 ± 2.52	7.35 ± 0.39	11.75 ± 1.73
GBR	0.87 ± 0.24	0.59 ± 0.21	7.08 ± 5.65	15.13 ± 3.91	5.50 ± 4.78	11.41 ± 3.41
ETR	0.91 ± 0.03	0.64 ± 0.11	7.37 ± 1.29	14.54 ± 2.22	5.63 ± 1.10	11.05 ± 1.63

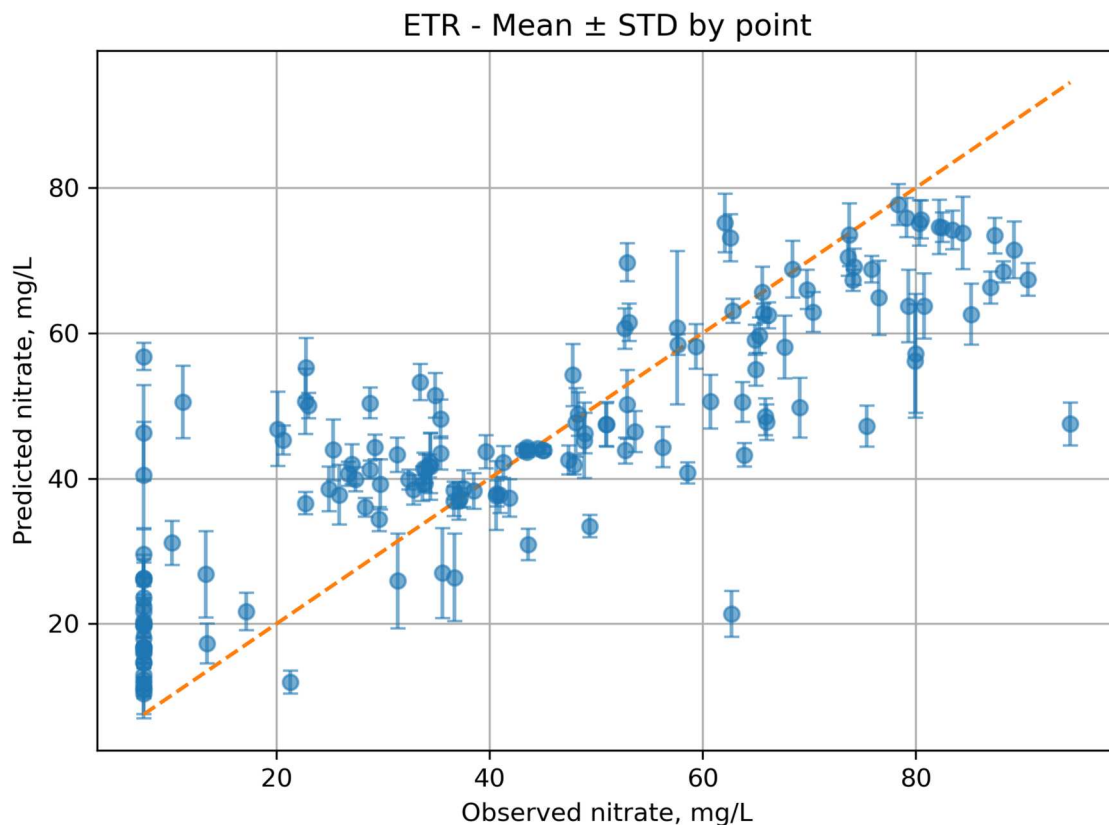
Compared to the classification models, the regression models are less accurate, and there is a larger gap between the training and testing metrics. This is to be expected, since the classification and regression problems are fundamentally different. The values for the regression models are in line with results from similar studies, such as Knoll *et al.* (2019), Liang *et al.* (2024), Mahboobi *et al.* (2023), and Ransom *et al.* (2022), and the metrics for the testing period are generally better (Table 4).

Figure 5 compares observed nitrate concentrations with the mean predicted values obtained from the ExtraTrees Regressor using out-of-sample predictions from the nested cross-validation procedure. The error bars represent the standard deviation of predictions across outer-loop validation iterations, providing an indication of prediction variability for each borehole. The model shows a tendency to underestimate the highest nitrate concentrations and to slightly

overestimate some of the lower observed values, indicating a regression-to-the-mean behaviour commonly observed in tree-based regression models. A clear cluster of boreholes is observed at the lower end of the concentration range, where nitrate values are constrained by the detection limit of the sampling method (11 boreholes with observed median nitrate concentrations of approximately 7.5 mg/L). This artificial concentration of observations at low nitrate levels may contribute to the reduced ability of the model to reproduce extreme high nitrate values.

3.3. Spatial predictions

Figure 6 presents the spatial distribution of nitrate obtained from classification- and regression-based modelling. The two methodologies provide complementary perspectives: while the classification model estimates the probability of exceeding the regulatory

**Figure 5.** Observed versus predicted nitrate concentrations (mg/L) for the ExtraTrees Regressor based on out-of-sample predictions from the nested cross-validation procedure. Each point represents an individual borehole, where the marker indicates the mean predicted value across outer cross-validation iterations and the error bars represent the standard deviation of predictions. The dashed line represents the 1:1 identity relationship ($y = x$).

threshold of 50 mg/L of nitrate, the regression model predicts continuous concentration levels across the study area.

In the classification map (Figure 6(A)), areas highlighted in red represent zones with a higher probability of exceeding the threshold, whereas blue areas indicate lower probabilities. The highest probabilities are primarily concentrated along the western boundary of the Chalk aquifer, which are regions where the aquifer is either unconfined or buried under thin confining deposits. Isolated zones of elevated probability are observed northeast of the city of Norwich and in the south-eastern edge of the aquifer. These spatial patterns are consistent with the existing conceptual model of nitrate contamination in the area, reinforcing the reliability of the classification-based predictions (Wang *et al.* 2016).

The predictive map produced by the regression model shows similar patterns (Figure 6(B)). High concentrations are predicted along the western boundary of the Chalk aquifer and in the surroundings of Norwich. Medium to high values of nitrate are also predicted along the north-eastern edge of the aquifer in the northern part and to the northwest of Ipswich. These patterns are more evident than in the classifier map.

Figure 6 also presents the predicted spatial distribution of nitrate contamination with the boundaries of the official Nitrate Vulnerable Zones (NVZ). Areas identified by both approaches as problematic exhibit general overlap with the extent of the NVZs. This is particularly evident in the Lincolnshire Chalk, located in the northern sub-region of the study area, and in the Anglian Chalk NVZ, along the western end of the southern sub-region. Figure S1 illustrates the spatial distribution of prediction uncertainty derived from the supervised learning models. Overall, higher variability is observed in the classification approach compared to regression, likely due to the inclusion of four different algorithms in the ensemble, whereas the regression results are based on a single model. In the regression case, moderate deviations are observed, with maximum values reaching approximately 10 mg/L, particularly in the southern part of the northern sector and to the east of Norwich. In contrast, the classification results show higher variability, with standard deviations of up to 0.15 in predicted probability, mainly in the western area where the Chalk outcrops and also to the east of Norwich. This pattern can be explained by differences among models, with some assigning probabilities close to 0.90 while others estimate values around 0.70 in these areas. Despite this variability, all models consistently identify these zones as the most problematic in terms of nitrate contamination.

Table 5 provides a detailed quantification of the results per surface area. It shows the areas predicted

to exceed 50 mg/L in classification, as well as the mean and median predicted nitrate concentrations for the regression approach, presented both for the entire study area and separately for NVZs. As shown, the study area covers 13,425 km², of which 9,477 km² are delineated as NVZs. For the classification-based approach, the average probability of exceeding the target threshold is 0.29, with a total of 1,969 km² (14.67% of the study area) exhibiting a probability greater than 50% of exceeding 50 mg/L. Within NVZs, 18.69% of the area (approximately 1,770 km²) shows a probability greater than 0.5, whereas in areas outside NVZs, only 5% of the area (198 km²) exceeds this threshold. In the regression-based approach, the predicted nitrate concentrations display an average value of 29.6 mg/L and a median of 23.8 mg/L across the study area, with more than 1,540 km² (11.4% of the total area) predicted to exceed 50 mg/L. Specifically, 15.6% of the NVZ area (1,476 km²) and 1.58% of the non-NVZ area (62 km²) are predicted to exceed the threshold, and both mean and median predicted concentrations are higher within NVZs than outside.

3.4. Influence of explanatory variables by means of Shapley additive explanation analysis

Figure 7 shows SHAP's bee swarm plot for the GBC for all the borehole locations. The most influential predictors are Chalk aquifer base elevation, thickness of superficial deposits, topographic slope, coarse fragment content, average temperature, distance to main statutory watercourses, topographic elevation and groundwater absolute level. The distribution of SHAP values not only shows the magnitude of each variable's contribution, but also illustrates how its intrinsic values affect the model's outcome. For instance, lower quaternary materials thickness values and higher Chalk aquifer base depth values have the same impact: a greater tendency toward contamination. Conversely, lower elevation values and wells closer to rivers exhibit a greater impact on nitrate contamination. Regarding variables representing possible sources of contamination, distance to agricultural areas and density of CDP were not found among the most important variables. However, it has not been observed that areas with a higher density of authorised discharge points are more likely to be classified as contaminated. Instead, a shorter distance from agricultural areas clearly has a greater impact on the model's results when classifying an area as contaminated (Figure S1).

Figure 8 shows the SHAP plot for the GBR model. As in the classification approach, the thickness of superficial deposits, Chalk aquifer base elevation, mean rainfall and groundwater level are significant features and show

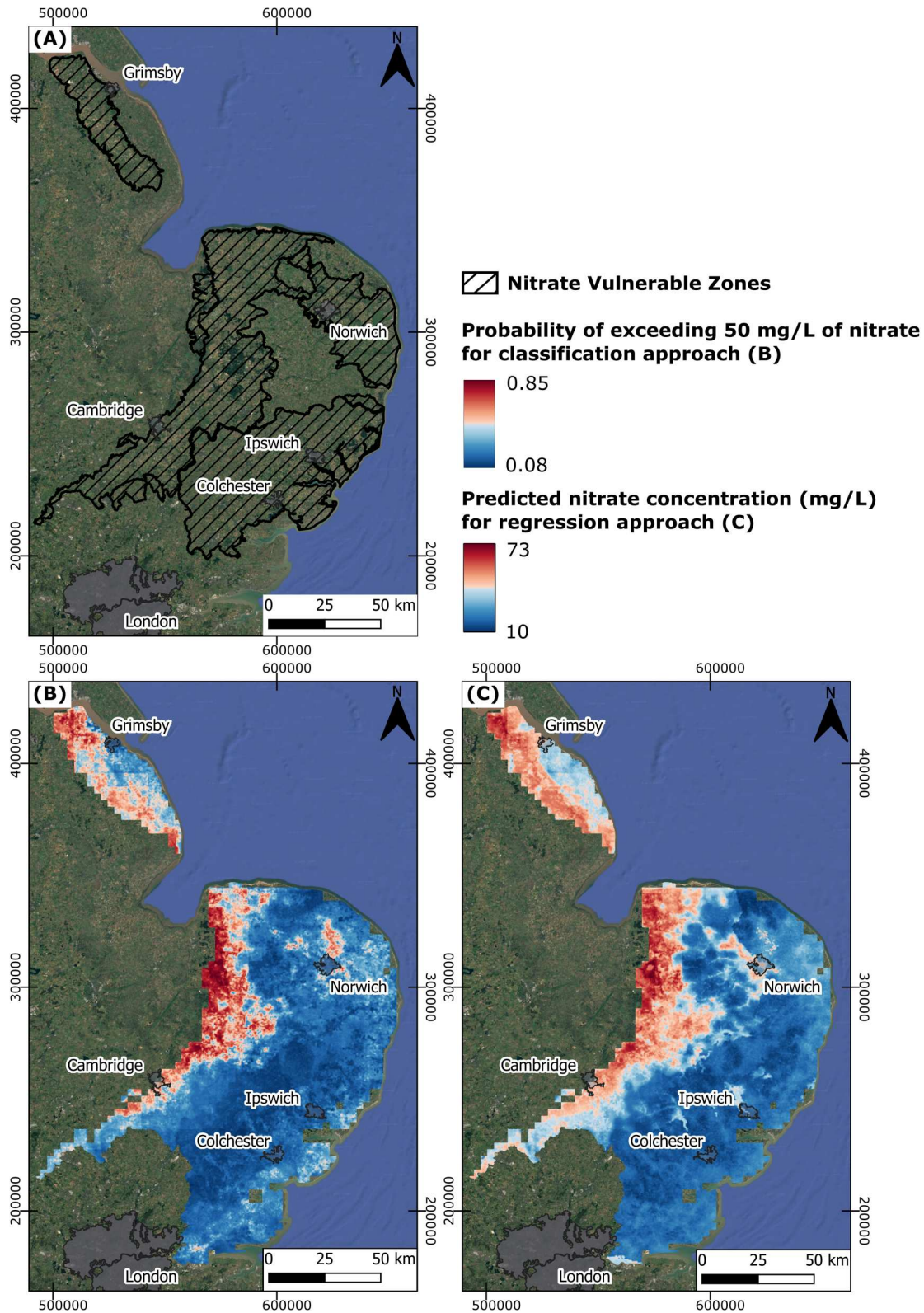


Figure 6. Predicted spatial distribution of nitrate contamination by means of supervised learning approaches: (A) Spatial designation of the Nitrate Vulnerable Zones (NVZ) (B) Probability of exceeding the 50 mg/L threshold according to the ensemble-based classification approach, and (C) Predictive map of nitrate concentration (mg/L) based on the ExtraTrees Regressor.

the same impacts (e.g. lower quaternary materials thickness gives higher predicted nitrate values).

Unlike GBC, however, ETR includes the ALC_{GRADE} factors as a potential nitrogen sources among the most influential predictor. The ALC_{GRADE} variable reflects the agricultural suitability of the land, with lower

values corresponding to higher-quality and more intensively managed agricultural areas. SHAP results indicate that lower grades (e.g. grades 2 and 3) have a positive contribution to the model output, suggesting an association with higher nitrate concentrations. In contrast, higher values corresponding to non-agricultural classes (e.g.

Table 5. Areas exceeding or remaining below probability thresholds for the classification approach, and areas predicted to exceed 50 mg/L with mean and median values for the regression approach, shown for the entire study area, nitrate vulnerable zones (NVZ), and non-designated areas.

Classification approach		
	Area (km ²)	% Area
Total – (All study area)	13425.25	100.00
Probability of exceeding the threshold of the target variable ≥ 0.5 – (All study area)	1969.25	14.67
Total (Inside NVZ)	9477.75	100
Probability of exceeding the threshold of the target variable ≥ 0.5 – (Inside NVZ)	1771.5	18.69
Total (Outside NVZ)	3947.5	100
Probability of exceeding the threshold of the target variable ≥ 0.5 – (Outside NVZ)	197.75	5.01
Regression approach		
Total – (All study area)	13425.25	100
Above 50 mg/L – (All study area)	1538.5	11.46
Average (mg/L) – (All study area)	29.61	
Median (mg/L) – (All study area)	23.85	
Total (Inside NVZ)	9477.75	100
Above 50 mg/L (Inside NVZ)	1476.25	15.58
Average (mg/L) – (Inside NVZ)	31.48	
Median (mg/L) – (Inside NVZ)	26.25	
Total (Outside NVZ)	3947.5	100
Above 50 mg/L (Outside NVZ)	62.25	1.58
Average (mg/L) – (Outside NVZ)		
Median (mg/L) – (Outside NVZ)	21.26	

class 6) show a negative contribution, indicating lower nitrate levels, which is consistent with reduced agricultural activity and associated nitrogen inputs (Figure 9). The density of CDP and distance to agriculture shows the same effect as in the classification approach, with no clear association between higher densities of authorised

discharge points and increased nitrate concentrations. Instead, proximity to agricultural areas remains the dominant factor influencing model predictions (Figure S2).

4. Discussion

4.1. Machine learning-based spatial predictions in the context of previous studies

Both the classification and regression approaches were based on data from the last four years (2020–2024), representing the most recent conditions and generally the most pessimistic scenario in terms of nitrate risk. Based on historical trends, Stuart and Kinniburgh (2005) predicted future nitrate concentrations at several Chalk monitoring sites for the period 2002–2017, revealing a general upward tendency in nitrate levels. This is consistent with the behaviour observed in several boreholes in our database.

Our outcomes are generally consistent with the existing knowledge. In particular, the spatial distribution of nitrate reported by Stuart and Kinniburgh (2005) is in agreement with the patterns obtained by classification and regression algorithms, which highlight the western outcrops of the Chalk aquifer as the most problematic areas. Our results also support those obtained by Heathcote (2024), who explain that the sector located 25 km north of Norwich, where their control boreholes are located and where the Chalk formation is covered by Pleistocene

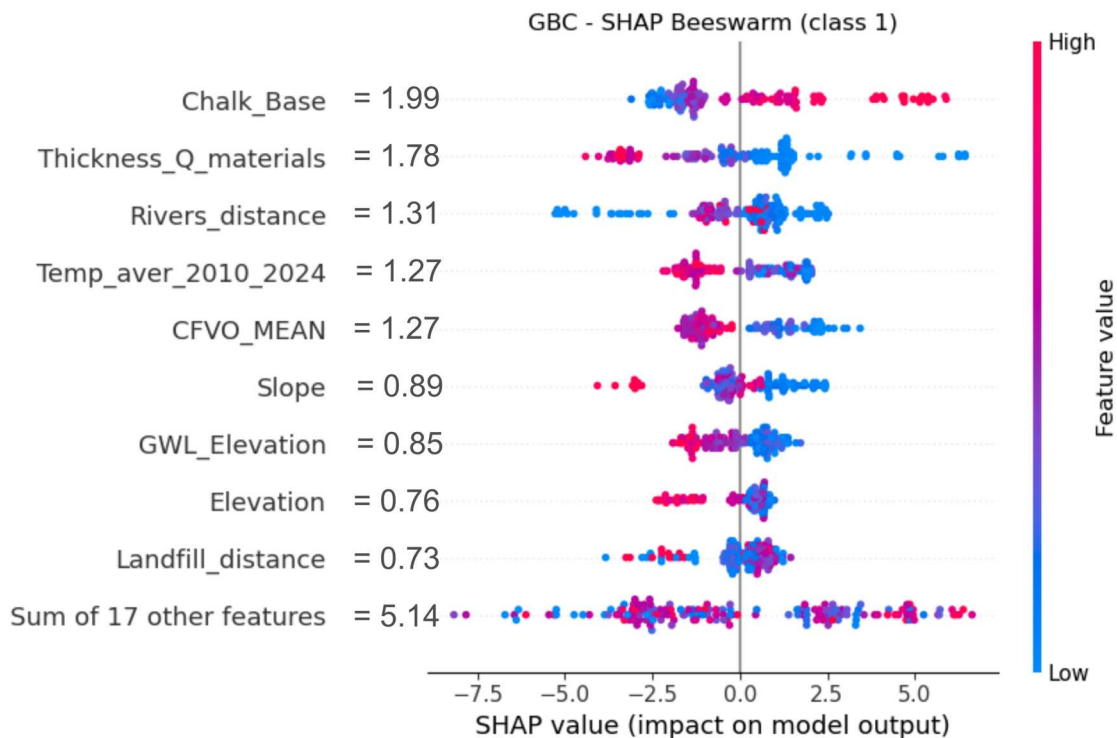


Figure 7. SHAP plot for the Gradient Boosting Classifier. The numbers displayed next to each predictor correspond to their mean absolute SHAP values, which determine the ranking of variable importance. The plot illustrates both the relative influence of each predictor and how their intrinsic values contribute to increasing or decreasing the predicted probability of exceeding the nitrate threshold.

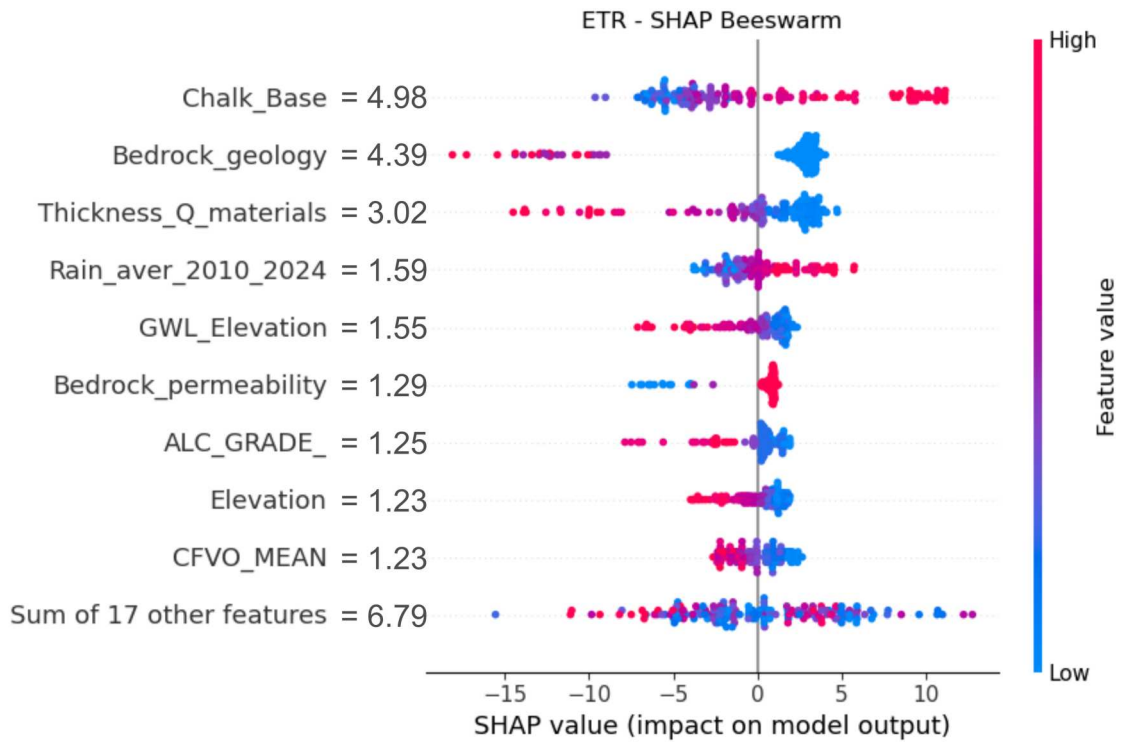


Figure 8. SHAP plot for the ExtraTrees Regressor. The numbers displayed next to each predictor correspond to their mean absolute SHAP values, which determine the ranking of variable importance. The plot illustrates both the relative influence of each predictor and how their intrinsic values contribute to increasing or decreasing the predicted nitrate value.

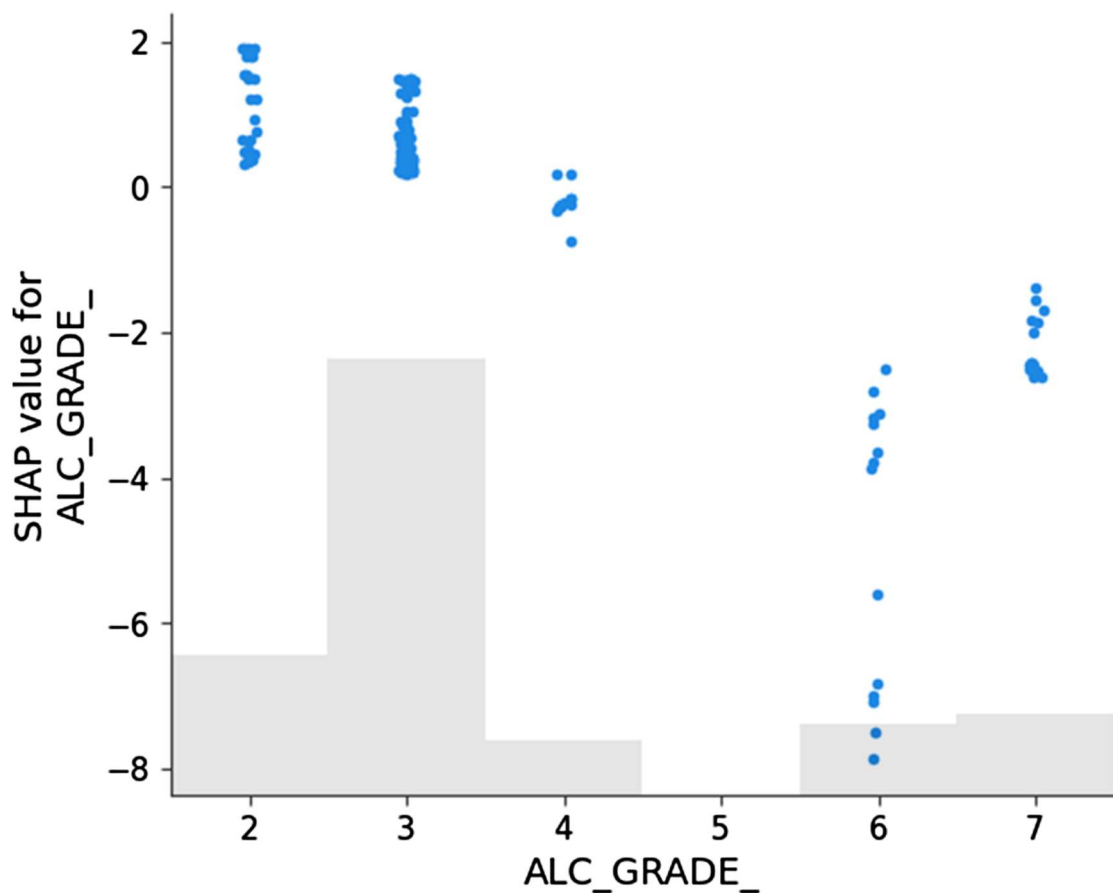


Figure 9. SHAP dependence plot for ALC_GRADE for the ExtraTrees Regressor (ETR) algorithm. The x-axis represents feature values and the y-axis their corresponding SHAP values. Grey bars along the x-axis indicate the distribution of observations.

deposits, remains largely nitrate-free. In the southern sector of the study area, around Ipswich and north of Colchester, this study reports a transition towards nitrate-bearing groundwater at several sites that can also be considered compatible with our maps.

Stuart *et al.* (2016) conclude that the Chalk aquifer dominates the national storage of legacy nitrate in the unsaturated zone, with particularly large historical masses in parts of East Anglia. The combination of thick unsaturated zones and slow travel times leads to substantial time lags, so current groundwater concentrations often reflect nitrate inputs from decades past. This may explain why some areas present high observed or predicted nitrate values despite moderate current surface pressures. These findings, consistent with those by Wang *et al.* (2012), highlight that historical agricultural inputs and unsaturated zone thickness are key factors controlling present-day nitrate contamination. Therefore, groundwater nitrate levels in thick unsaturated areas may not yet reflect recent reductions in loading. Consistently, SHAP analysis indicates that the most vulnerable zones identified by both the classification and regression approaches are strongly influenced by geological controls, rather than solely by current surface inputs.

The two modelling approaches produce broadly consistent spatial patterns of nitrate occurrence in groundwater. Both highlight the western boundary of the Chalk aquifer and the north-eastern area of Norwich as areas of high contamination. Some differences emerge when comparing the predictive structures of the two approaches. The classification model generates a more discrete spatial pattern, providing a clear distinction between areas with a high and low probability of exceeding the regulatory threshold. In contrast, the regression model provides a continuous gradient of predicted nitrate concentrations, offering a more nuanced representation of areas of intermediate susceptibility. This enables us to identify the strengths and weaknesses of both approaches. The classification approach can discern areas with high and low nitrate contamination potential with great precision, but intermediate probabilities do not provide relevant information to stakeholders, since a probability of 0.5 is virtually meaningless. Conversely, the regression approach may present greater uncertainty due to the type of target variable it predicts spatially. However, in intermediate areas, it enables us to obtain a median nitrate concentration value, which may be more useful to stakeholders in terms of decision-making.

A key difference between the two approaches emerges in the southern coastal area, where the regression model predicts higher nitrate concentrations than the classification map indicates. This is probably because the regression model can capture subtle variations in the underlying predictors that

are not sufficient to shift probabilities beyond the 0.5 threshold in the classification approach. Overall, the regression model provides a more nuanced depiction of nitrate contamination, with low values in central areas, increasing towards the coast and slightly towards the west, and high values in the unconfined Chalk zone and the south-eastern sector. The classification model, by its very nature, provides a more categorical delineation of contamination-prone areas.

4.2. Implications of Shapley additive explanation analysis

SHAP analyses highlight the role of geological factors as predictors for nitrate contamination. This control is particularly evident in the western part of the Chalk aquifer, where unconfined conditions favour nitrate leaching. By contrast, the eastern part of the aquifer is characterised by Quaternary deposits that confine the aquifer and restrict vertical infiltration.

The agricultural zones also emerge as an important explanatory variable, which is consistent with the conceptual model of nitrate sources and transport in the study area. However, a reverse trend is observed in the case of consented discharge points (CDPs): areas with a higher density of CDPs are not necessarily associated with higher nitrate pollution. This finding highlights the importance of model interpretability and is likely due to the concentration of CDPs in urban areas, whereas higher nitrate levels are primarily found in agricultural areas. Furthermore, using CDP locations without information on actual discharge loads limits the model's ability to capture their real impact, suggesting a spatial association rather than direct causality.

The high importance of edaphic factors (soil coarse fragment) has been observed in both approaches. These results are consistent with those obtained in similar studies conducted in other regions. Karimanzira *et al.* (2023) found that the most important predictors for nitrate were soil characteristics, geological features, and proximity to rivers, while Ransom *et al.* (2022) showed that the most important variables were well characteristics, soil types, climate and hydrology variables, and land use. Serra *et al.* (2024) attributed importance to soil-related factors (bulk density and soil organic carbon), hydrogeological factors (aquifer thickness) and land use factors (rainfed crops). Knoll *et al.* (2019) identified hydrogeological units, the percentage of arable land, and nitrogen balance as the most relevant predictors. Hiscock *et al.* (2024) demonstrated that the presence of low-permeability glacial tills significantly limits the penetration of agricultural nitrate (NO_3^-) within the Blackwater Drain sub-catchment of the River Wensum. In contrast, in areas dominated by higher-

permeability glacial sands and gravels, nitrate is able to penetrate the superficial deposits more easily.

As these geological materials are present in the study area, our findings suggest that the permeability of superficial deposits is an important factor in controlling nitrate infiltration into groundwater. This is consistent with the findings of Wang *et al.* (2013), who showed that nitrate reaches groundwater more rapidly where the unsaturated zone is thin, whereas in areas with thick unsaturated zones, peak concentrations may be significantly delayed. These authors identified unsaturated zone thickness as the main factor controlling nitrate behaviour in the Chalk and Permo-Triassic sandstone aquifers. Together, these findings reinforce the dominant role of geological factors in governing nitrate transfer to groundwater in the Chalk aquifer.

4.3. Management implications

Our findings provide an independent comparison benchmark for the official delineation of NVZs. In this context, both modelling approaches suggest that there are areas outside the perimeter of vulnerable zones that show similar markers to areas known to be contaminated. Using the classification approach, 15% of the total area has a probability greater than 0.5 of exceeding 50 mg/L, of which 5% (approximately 200 km²) is located outside NVZs. Similarly, the regression approach predicts that 11.5% of the study area exceeds the 50 mg/L threshold, with 1.5% (62 km²) lying outside designated vulnerable zones. In both approaches, the mean and median predicted nitrate values, and the predicted probabilities, are higher within NVZs, confirming that these areas are generally more affected. From a broader perspective, all this appears consistent with a recent European assessment (Serra *et al.* 2024), which reported that a significant proportion of nitrate hotspots occur outside officially designated NVZs. This suggests that machine learning approaches such as the one at hand could provide added value in terms of enhancing the delineation of nitrate-affected areas.

Water companies produce drinking water safety plans for raw water sources, which include risks to the catchment raw water quality. The plans often consider the sources and pathways of potential contaminants in the catchment and the risk is validated with water quality samples. This assessment could help indicate areas where new sources are most prone to nitrate contamination. One challenge with nitrate is that it is possible to have a similar land use in a locality, but, due to the influence of the denitrification processes as a result of the soil types and superficial geology, obtain different nitrate concentrations in the underlying aquifer. This assessment can help indicate areas where the nitrate concentration may be

lower using readily available datasets. The bee swarm SHAP plots seem to align well with known catchment processes. The quaternary deposits in the area vary, with some significant thicknesses of lower permeability glacial deposits which impede the vertical migration of water and allow for greater denitrification. In this context, one interesting factor was the fact that the distance to agriculture variable did not influence the overall result as greatly as the physical catchment property variables, even though agriculture is considered the primary source of the aquifer nitrate contamination.

Therefore, when considering new borehole placements in areas where nitrate concentration is a concern, areas with a greater thickness of quaternary deposits may be preferable. However, where the aquifer is covered with large thicknesses of quaternary deposits, there may be other geo-chemical process of concern to the quality of drinking water. Furthermore, potential yield of the borehole and the influence radius are important placement considerations alongside the regulator's licensing policy dictating where water is available for abstraction. The assessment provides potentially useful insight into the possible nitrate concentration within the Chalk aquifer, which amongst other factors can assist in borehole placement and potential raw water treatment decision making.

Finally, the insights provided by the application of machine learning techniques can contribute to evaluate future risks associated with legacy nitrate contamination, as influenced by the various natural factors as well as likely historical land use—which has not significantly changed over the past few decades—and perhaps informed by existing or potential data on nitrate levels at various depths. This could in turn be used to refine recharge estimates and groundwater models, including identifying areas where greater geographical resolution would be beneficial.

4.4. Classification vs regression: advantages and limitations

A key distinction between the two approaches lies in the type of information they provide. The classification model estimates the probability of exceeding a regulatory threshold, making it particularly well suited for risk assessment, stakeholder communication, and decision-making. By contrast, the regression model predicts continuous nitrate concentrations. This is especially informative in areas where exceedance risk is high but predicted values lie close to the threshold, as it indicates the magnitude of intervention required and can help anticipate treatment needs.

In terms of performance, classification metrics were consistently strong across multiple algorithms, suggesting that the exceedance problem is relatively

tractable and supporting a high degree of confidence in the outputs. This consistency also enables the use of multi-model blending to produce more robust predictions. For regression, performance varied more substantially between algorithms, which reduces confidence in individual model outputs and makes ensemble approaches less straightforward to justify.

Regarding adaptability, both approaches are transferable to other regions and contaminants. Classification offers a natural framework for spatial risk mapping, while regression may be more readily extended to spatio-temporal settings where predicting continuous changes is important. However, in this study both approaches represent a static snapshot of the 2020–2024 period, as temporal dynamics are not explicitly modelled. As a result, neither framework captures temporal trends, emerging patterns, or intra-annual variability, which may limit their usefulness for fully dynamic groundwater management. There are benefits to using both approaches simultaneously. Firstly, they provide complementary information so can be considered together to assess how likely a region is to be contaminated. Secondly, since they make use of different subsets of explanatory variables, a deeper level of process understanding can be derived using interpretable machine learning techniques such as SHAP analysis. However, if limited to one approach, data requirements could constrain the choice between classification and regression. If a probability of exceedance is required, then the classification approach is most suitable, and if the level of contamination is of interest then the regression approach is best.

Our approach uses tree-based ensemble algorithms were implemented only. This choice reflects their well-established robustness and strong performance in spatial prediction tasks involving hydrochemical variables and geoenvironmental factors, as documented in previous studies (Knoll *et al.* 2019, Ransom *et al.* 2022, Karimanzira *et al.* 2023, Covatti *et al.* 2025). However, this focus necessarily narrows the methodological scope. Alternative approaches – such as artificial neural networks or linear models (e.g. logistic regression) – have shown promise in related applications and could offer complementary perspectives or potential gains in predictive performance (Nafouanti *et al.* 2021, Karimanzira *et al.* 2023, Gholami *et al.* 2024). Hyperparameter optimisation was conducted using a random search cross-validation strategy, which provides a practical balance between performance and computational cost. While this approach is widely adopted, it does not guarantee identification of globally optimal configurations. Future work could therefore explore more advanced optimisation techniques, including metaheuristic or evolutionary algorithms, to further refine model calibration and enhance predictive accuracy, as suggested in previous studies (Moayedi *et al.* 2020).

4.5. Data limitations

A significant effort was made to collate the available data for the target and explanatory variables. However, there were limitations in time and data availability which resulted in compromises. Firstly, our modelling process would be improved by including water quality data from other abstractors also tapping the Chalk aquifers. There is also potential to explore the utility of different time-lags of mean rainfall and temperature, along with other variables which may vary with time (e.g. land use). This avenue was not explored further, partly due to time constraints but also because any benefit of including additional variables may be outweighed by the addition of noise to the data. Experience from other ongoing research suggests that such complexity does not always lead to substantial improvements in predictive performance. Furthermore, most existing studies in the literature rely on long-term average values of climatic variables rather than time-lagged representations, which supports our choice (Knoll *et al.* 2019, Ransom *et al.* 2022, Karimanzira *et al.* 2023, Serra *et al.* 2024). Additionally, although some studies include hydrochemical variables as predictors (Deng *et al.* 2023, Abba *et al.* 2024), this study did not consider their use due to a lack of spatial continuity and associated uncertainty in interpolation. Other works (Xenakis *et al.* 2026) point out the possibility of employing both a geospatial approach and an in-situ approach, with the latter potentially including hydrochemical variables; however, this is not feasible if the goal is to produce spatial mapping, which is one of the main objectives of this study. Furthermore, since certain hydrochemical parameters are closely linked to the processes controlling nitrate concentrations, including them could introduce redundancy or indirect information related to the target variable. This could lead to overly optimistic model performance.

We used the most recent land-use information, in line with the literature (Knoll *et al.* 2019, Karimanzira *et al.* 2023, Mahlknecht *et al.* 2023). Although present-day nitrate concentrations may reflect historical inputs, rather than recent contamination, we were unable to find a dataset of past land uses. Furthermore, incorporating multiple land-use layers would have introduced strong multicollinearity issues that would have required careful consideration. On a related note, the Consented Discharge Points (CDP) dataset only provides the geographic location of discharge sites and lacks data on concentration, load volumes or temporal behaviour. This constrains the capacity of the models to quantify the actual impact of these discharges on groundwater nitrate levels, which may partly explain the relatively low importance of the CDP dataset in the SHAP analysis. Furthermore, CPDs in England tend to be urban focused, which

implies that other contaminants, such as phosphorous, may be more dominant than nitrate.

The modelling workflow and code implementation do not require predictors to be in the same spatial resolution. However, heterogeneity introduces constraints when analysing the results. In particular, the most influential predictors, such as geological and soil-related variables, have native resolutions generally ranging from 250 m to 1 km. Therefore, it would be unrealistic to expect the predictive maps to achieve a finer level of spatial precision than that provided by the underlying data. For this reason, the final predictive surface was generated at a resolution of 500×500 m, representing a balance between computational efficiency and the informational content of the input datasets.

5. Conclusions

From a site-specific perspective, this study demonstrates that data-driven spatial modelling, when embedded within a robust conceptual framework of aquifer contamination, can successfully reproduce known patterns of nitrate occurrence in groundwater. Both classification and regression approaches identify the unconfined parts of the Chalk aquifer as the most vulnerable to nitrate contamination, which is consistent with long-term monitoring evidence and previous regional studies. Results also highlight the dominant role of geological and hydrogeological controls – such as aquifer confinement, unsaturated zone thickness and superficial deposits – in regulating nitrate transfer from the surface to groundwater.

A comparison of modelling approaches shows that classification and regression provide complementary interpretations to the existing datasets. Classification models are well suited for regulatory and risk-based contexts, offering clear probabilities of exceeding statutory thresholds and performing consistently well across multiple algorithms. Regression-based results reflect the greater difficulty of achieving precise predictive performance for continuous outcomes within the limitations and structure of the available dataset. Regression models however provide a continuous representation of nitrate concentrations that is valuable for understanding intermediate-risk areas and informing treatment and operational decisions. Differences between the approaches, particularly in the southern coastal zone of the study area, underscore how continuous models can capture subtle spatial gradients that may not translate into threshold exceedance probabilities. Together, the two approaches enhance process understanding and decision support, especially when combined with interpretable machine learning tools such as SHAP to elucidate the relative importance of contamination predictors.

From a management perspective, the findings suggest that the official delineation of nitrate vulnerable

zones generally corresponds to those regions with higher predicted nitrate levels. However, both modelling approaches also point to areas outside the designated zones where high nitrate concentrations may occur, indicating potential blind spots in current monitoring strategies. In this context, our results provide a practical, data-driven basis for prioritising the siting of new monitoring boreholes and refining existing surveillance networks. Importantly, this study demonstrates the added value of integrating machine learning with stakeholder knowledge and operational datasets – an approach that remains relatively uncommon in the literature but is critical for capturing real-world system complexity. Although our analysis is subject to data limitations – particularly regarding borehole density, discharge load information, and temporal dynamics – the framework is transparent, adaptable, and readily transferable to other regions. As such, it offers a robust foundation for improving groundwater risk assessment and supporting more targeted, evidence-based management decisions.

Acknowledgements

The authors would like to thank the British Geological Survey for kindly making available the geological maps of the study region.

Author contributions

CRedit: **V. Gómez-Escalonilla:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Writing – original draft; **H. Baron:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing; **T. Read:** Conceptualization, Formal analysis, Methodology, Supervision, Validation, Writing – original draft, Writing – review & editing; **J. Watson:** Conceptualization, Formal analysis, Methodology, Validation, Writing – original draft; **M. Rodríguez Del Rosario:** Data curation, Investigation; **G. Darch:** Conceptualization, Resources, Supervision; **V. Keller:** Conceptualization, Formal analysis, Writing – review & editing; **N. Rickards:** Conceptualization, Formal analysis, Writing – review & editing; **P. Martínez-Santos:** Conceptualization, Formal analysis, Writing – review & editing.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work has been funded under the Stars4Water project (Research grant 101059372 of the European Commission, HORIZON-CL6-2021-CLIMATE-01).

ORCID

H. Baron  <http://orcid.org/0000-0003-0070-8247>

References

- Abascal, E., *et al.*, 2022. Global diagnosis of nitrate pollution in groundwater and review of removal technologies. *Science of the Total Environment*, 810 (1), 152223.
- Abba, S.I., *et al.*, 2024. Nitrate concentrations tracking from multi-aquifer groundwater vulnerability zones: insight from machine learning and spatial mapping. *Process Safety and Environmental Protection*, 184, 1143–1157. doi:10.1016/j.psep.2024.02.041.
- Asadi Nalivan, O., *et al.*, 2023. A hybrid intelligent model for spatial analysis of groundwater potential around Urmia Lake, Iran. *Stochastic Environmental Research and Risk Assessment*, 37 (5), 1821–1838.
- BGS, 2016a. BGS geology 1:50 000 (DiGMapGB-50) superficial maps (V8). British Geological Survey. BGS Digital Data under Licence No. 2023/030A & 2023/030B British Geological Survey© and Database Right UKRI. All rights reserved.
- BGS, 2016b. Superficial deposit thickness (1Km HexGrid) dataset. British Geological Survey. BGS Digital Data under Licence No. 2023/030A & 2023/030B British Geological Survey© and Database Right UKRI. All rights reserved.
- BGS, 2021a. Permeability 1:50 000. British Geological Survey. BGS Digital Data under Licence No. 2023/030A & 2023/030B British Geological Survey© and Database Right UKRI. All rights reserved.
- BGS, 2021b. Soil parent material dataset 1:50 000. British Geological Survey. BGS Digital Data under Licence No. 2023/030A & 2023/030B British Geological Survey© and Database Right UKRI. All rights reserved.
- BGS, 2025. The Chalk. British Geological Survey. Available from: <https://www2.bgs.ac.uk/groundwater/shaleGas/aquifersAndShales/maps/aquifers/Chalk.html> [Accessed December 2025].
- Bhattarai, A., *et al.*, 2021. Prediction of nitrate and phosphorus concentrations using machine learning algorithms in watersheds with different landuse. *Water*, 13 (21), 3096. doi:10.3390/w13213096.
- Bilgilioglu, S.S., *et al.*, 2025. Explainable sinkhole susceptibility mapping using machine-learning-based SHAP: quantifying and comparing the effects of contributing factors in konya, türkiye. *Applied Sciences*, 15 (6), 3139. doi:10.3390/app15063139.
- Breiman, L., 1996. Bagging predictors. *Machine Learning*, 24, 123–140.
- Breiman, L., 2001. Random forests. *Machine Learning*, 45, 5–32.
- Cawley, G.C. and Talbot, N.L.C., 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11, 2079–2107.
- Chahid, M., *et al.*, 2025. Deep learning framework for mapping nitrate pollution in coastal aquifers under land use pressure. *Scientific Reports*, 15, 34946. doi:10.1038/s41598-025-18996-7.
- Covatti, G., *et al.*, 2025. Nitrate contamination in groundwater across Switzerland: spatial prediction and data-driven assessment of anthropogenic and environmental drivers. *Science of the Total Environment*, 973, 179121.
- Deng, Y., Ye, X., and Du, X., 2023. Predictive modeling and analysis of key drivers of groundwater nitrate pollution based on machine learning. *Journal of Hydrology*, 624, 129934.
- Díaz-Alcaide, S. and Martínez-Santos, P., 2019. Advances in groundwater potential mapping. *Hydrogeology Journal*, 27 (9), 2307–2324.
- Dormann, C.F., *et al.*, 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36 (1), 27–46.
- Duran-Llacer, I., *et al.*, 2025. Approach to mapping groundwater-dependent ecosystems through machine learning in central Chile. *Groundwater for Sustainable Development*, 31, 101526.
- Elsayed, A., *et al.*, 2025. Regression-based machine learning models for nitrate and chloride prediction in surface water in a small agricultural sand plain sub-watershed in southwestern Ontario, Canada. *Frontiers in Environmental Science*, 13, 1543852. doi:10.3389/fenvs.2025.1543852.
- Environment Agency, 2023. Statutory main river map. Available from: <https://environment.data.gov.uk/dataset/25dde009-ba7d-40de-8380-c5c3bb32ccdc>.
- Eryigit, M. and Engel, B., 2022. Spatiotemporal modelling of groundwater flow and nitrate contamination in an agriculture-dominated watershed. *Journal of Environmental Informatics*. doi:10.3808/jei.202100470.
- Freund, Y. and Schapire, R.E., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55, 119–139.
- Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 29 (5), 1189–1232.
- Gaines, T.P. and Gaines, S.T., 1994. Soil texture effect on nitrate leaching in soil percolates. *Communications in Soil Science and Plant Analysis*, 25 (13–14), 2561–2570.
- García-Aróstegui, J.L., Baudron, P., and Robles-Arenas, V.M., 2024. Sampling methods may drive short-term groundwater nitrate variability in an irrigated watershed connected to a coastal lagoon (Campo de Cartagena-Mar Menor, SE Spain). *Science of The Total Environment*, 912, 169188, doi:10.1016/j.scitotenv.2023.169188.
- Géron, A., 2017. *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly Media, Inc.
- Geurts, P., Ernst, D., and Wehenkel, L., 2006. Extremely randomized trees. *Machine Learning*, 63, 3–42.
- Gholami, V., *et al.*, 2024. Combining artificial neural networks and genetic algorithms to model nitrate contamination in groundwater. *Natural Hazards*, 120 (5), 4789–4809.
- Gómez-Escalonilla, V., *et al.*, 2022b. Delineation of groundwater potential zones by means of ensemble tree supervised classification methods in the Eastern Lake Chad basin. *Geocarto International*, 37 (25), 8924–8951.
- Gómez-Escalonilla, V., *et al.*, 2024. A machine learning approach to site groundwater contamination monitoring wells. *Applied Water Science*, 14 (12), 250. doi:10.1007/s13201-024-02320-1.
- Gómez-Escalonilla, V., Martínez-Santos, P., and Martín-Loeches, M., 2022a. Preprocessing approaches in machine-learning-based groundwater potential mapping: an application to the Koulikoro and Bamako regions, Mali. *Hydrology and Earth System Sciences*, 26 (2), 221–243.
- He, S., *et al.*, 2022. Predictive modeling of groundwater nitrate pollution and evaluating its main impact factors using random forest. *Chemosphere*, 290, 133388. doi:10.1016/j.chemosphere.2021.133388.

- Heathcote, J.A., 2024. Long-term changes in Chalk groundwater quality in East Anglia, UK. *Quarterly Journal of Engineering Geology and Hydrogeology*, 58 (1), qjehg2024-072. doi:10.1144/qjehg2024-072.
- Heung, B., Bulmer, C.E., and Schmidt, M.G., 2014. Predictive soil parent material mapping at a regional-scale: a random forest approach. *Geoderma*, 214, 141–154.
- Hiscock, K.M., et al., 2024. Influence of Pleistocene glacial deposits on the transport of agricultural nitrate in the river Wensum catchment, UK. *Journal of Hydrology*, 633, 130982. doi:10.1016/j.jhydrol.2024.130982.
- Jiang, Y. and Somers, G., 2009. Modeling effects of nitrate from non-point sources on groundwater quality in an agricultural watershed in Prince Edward Island, Canada. *Hydrogeology Journal*, 17, 707–724. doi:10.1007/s10040-008-0390-2.
- Karimanzira, D., et al., 2023. Application of machine learning and deep neural networks for spatial prediction of groundwater nitrate concentration to improve land use management practices. *Frontiers in Water*, 5, 1193142.
- Kearsey, T., Lee, J., and Gow, H., 2019. User guide for the British geological survey buried valleys (onshore) version 1.
- Knoll, L., Breuer, L., and Bach, M., 2019. Large scale prediction of groundwater nitrate concentrations from spatial data using machine learning. *Science of The Total Environment*, 668, 1317–1327.
- Krstajić, D., et al., 2014. Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of Cheminformatics*, 6, 10. doi:10.1186/1758-2946-6-10.
- Lee, J.M., Ko, K.S., and Yoo, K., 2023. A machine learning-based approach to predict groundwater nitrate susceptibility using field measurements and hydrogeological variables in the Nonsan Stream Watershed, South Korea. *Applied Water Science*, 13, 242. doi:10.1007/s13201-023-02043-9.
- Lerner, D.N. and Harris, B., 2009. The relationship between land use and groundwater resources and quality. *Land Use Policy*, 26, S265–S273.
- Liang, Y., et al., 2024. Mapping specific groundwater nitrate concentrations from spatial data using machine learning: a case study of Chongqing, China. *Heliyon*, 10, 2405–8440. DOI: 10.1016/j.heliyon.2024.e27867.
- Lundberg, S.M., et al., 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2, 56–67. doi:10.1038/s42256-019-0138-9.
- Lundberg, S.M. and Lee, S.I., 2017. A unified approach to interpreting model predictions. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Red Hook, NY, USA: Curran Associates Inc, 4768–4777.
- Mahboobi, H., Shakiba, A., and Mirbagheri, B., 2023. Improving groundwater nitrate concentration prediction using local ensemble of machine learning models. *Journal of Environmental Management*, 345, 118782.
- Mahlknecht, J., et al., 2023. Nitrate prediction in groundwater of data scarce regions: the futuristic fresh-water management outlook. *Science of The Total Environment*, 905, 166863. doi:10.1016/j.scitotenv.2023.166863.
- Makumbura, R.K., et al., 2024. Advancing water quality assessment and prediction using machine learning models, coupled with explainable artificial intelligence (XAI) techniques like Shapley additive explanations (SHAP) for interpreting the black-box nature. *Results in Engineering*, 23, 102831, doi:10.1016/j.rineng.2024.102831.
- Martínez-Santos, P., et al., 2021. Predictive mapping of aquatic ecosystems by means of support vector machines and random forests: an application to the Valle del Cauca region, Colombia. *Journal of Hydrology*, 595 (2021), 126026. DOI: 10.1016/j.jhydrol.2021.126026.
- Martínez-Santos, P., Gómez-Escalonilla, V., Díaz-Alcaide, S., et al., 2025. A surrogate approach to model groundwater level in time and space based on tree regressors. *Applied Water Science*, 15, 206. doi:10.1007/s13201-025-02572-5.
- Mehrabi, M., Polya, D.A., and Han, Y., 2025. Machine learning models of the geospatial distribution of groundwater quality: a systematic review. *Water*, 17, 2861. doi:10.3390/w17192861.
- Met Office, et al., 2024. *HadUK-Grid gridded climate observations on a 1 km grid over the UK, v1.3.0.ceda (1836-2023)*. NERC EDS Centre for Environmental Data Analysis. 18 July 2024. doi:10.5285/b963ead70580451aa7455782224479d5.
- Moayed, H., et al., 2020. Fuzzy-metaheuristic ensembles for spatial assessment of forest fire susceptibility. *Journal of Environmental Management*, 260, 109867.
- Morton, R.D., et al., 2020. *Land cover map 2019 (20 m classified pixels, GB)*. NERC Environmental Information Data Centre.
- Mosavi, A., et al., 2021. Ensemble boosting and bagging based machine learning models for groundwater potential prediction. *Water Resources Management*, 35 (1), 23–37.
- Nafouanti, M.B., et al., 2021. Prediction on the fluoride contamination in groundwater at the Datong Basin, Northern China: comparison of random forest, logistic regression and artificial neural network. *Applied Geochemistry*, 132, 105054.
- Pedregosa, F., et al., 2011. Scikit-learn: machine learning in python. *Journal of Machine Learning Research*, 12 (2011), 2825–2830.
- Podgorski, J. and Berg, M., 2020. Global threat of arsenic in groundwater. *Science*, 368 (6493), 845–850.
- Poggio, L., et al., 2021. Soilgrids 2.0: producing soil information for the globe with quantified spatial uncertainty. *Soil*, 7 (1), 217–240.
- Ransom, K.M., et al., 2022. Machine learning predictions of nitrate in groundwater used for drinking supply in the conterminous United States. *Science of the Total Environment*, 807, 151065.
- Ribeiro, M.T., Singh, S., and Guestrin, C., 2016. Model-agnostic interpretability of machine learning. *arXiv Preprint arXiv:1606.05386*.
- Rivett, M.O., et al., 2007. Nitrate occurrence and attenuation in the major aquifers of England and Wales. *Quarterly Journal of Engineering Geology and Hydrogeology*, 40 (4), 335–352. doi:10.1144/1470-9236/07-032.
- Rodríguez-Galiano, V., et al., 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). *Science of the Total Environment*, 476, 189–206.
- Rosecrans, C.Z., et al., 2022. Predicting regional fluoride concentrations at public and domestic supply depths in basin-fill aquifers of the western United States using a random forest model. *Science of The Total Environment*, 806, 150960. doi:10.1016/j.scitotenv.2021.150960.

- Rosenstock, T.S., *et al.*, 2014. Agriculture's contribution to nitrate contamination of Californian groundwater (1945–2005). *Journal of Environmental Quality*, 43 (3), 895–907.
- Serra, J., *et al.*, 2024. Assessing nitrate groundwater hotspots in Europe reveals an inadequate designation of Nitrate Vulnerable Zones. *Chemosphere*, 355, 141830. doi:10.1016/j.chemosphere.2024.141830.
- Singh, S., *et al.*, 2022. Nitrates in the environment: a critical review of their distribution, sensing techniques, ecological effects and remediation. *Chemosphere*, 287, 131996.
- Spalding, R.F. and Exner, M.E., 1993. Occurrence of nitrate in groundwater—a review. *Journal of Environmental Quality*, 22, 392–402. DOI:10.2134/jeq1993.00472425002200030002x.
- Spijker, J., Fraters, D., and Vrijhoef, A., 2021. A machine learning based modelling framework to predict nitrate leaching from agricultural soils across the Netherlands. *Environmental Research Communications*, 3 (4), 045002.
- Stuart, M.E., *et al.*, 2016. Modelling the groundwater nitrate legacy. British Geological Survey, 75 (OR/16/036). <https://nora.nerc.ac.uk/id/eprint/516512/1/OR16036.pdf>.
- Stuart, M.E. and Kinniburgh, D.G., 2005. *Nitrate trends in groundwater*. Nottingham, UK: British Geological Survey Internal Report, 74 (IR/05/137). <https://nora.nerc.ac.uk/id/eprint/536076/1/IR05137.pdf>.
- Sun, H., *et al.*, 2024. Variation of nitrate sources affected by precipitation with different intensities in groundwater in the piedmont plain area of alluvial-pluvial fan. *Journal of Environmental Management*, 367, 121885.
- Taylor, P.J., 2021. *Spatial inventory of UK waterbodies*. NERC EDS Environmental Information Data Centre (Dataset). doi:10.5285/b6b92ce3-dcd7-4f0b-8e43-e937ddf1d4eb.
- Wang, J., *et al.*, 2023. Surface-Water nitrate exposure to world populations has expanded and intensified during 1970–2010. *Environmental Science & Technology*, 57 (48), 19395–19406. December 5. doi:10.1021/acs.est.3c06150. Epub 2023 Nov 21. PMID: 38050814; PMCID: PMC10702521.
- Wang, L., *et al.*, 2012. Prediction of the arrival of peak nitrate concentrations at the water table at the regional scale in Great Britain. *Hydrological Processes*, 26, 226–239. doi:10.1002/hyp.8164.
- Wang, L., *et al.*, 2013. The nitrate time bomb: a numerical way to investigate nitrate storage and lag time in the unsaturated zone. *Environmental Geochemistry and Health*, 35 (5), 667–681, October. doi:10.1007/s10653-013-9550-y. Epub 2013 Jun 26. PMID: 23801341.
- Wang, L., *et al.*, 2016. The changing trend in nitrate concentrations in major aquifers due to historical nitrate loading from agricultural land across England and Wales from 1925 to 2150. *Science of the Total Environment*, 542 (Part A), 694–705. doi:10.1016/j.scitotenv.2015.10.127.
- Xenakis, G.I., *et al.*, 2026. Controls and predictions of geogenic redox-sensitive contaminants in Danish groundwater. *Groundwater for Sustainable Development*, 33, 101600.
- Yang, H., *et al.*, 2023. Prediction of phosphorus concentrations in shallow groundwater in intensive agricultural regions based on machine learning. *Chemosphere*, 313, 137623.
- Yang, S., *et al.*, 2024. Spatial mapping and prediction of groundwater quality using ensemble learning models and Shapley additive explanations with spatial uncertainty analysis. *Water*, 16, 2375. doi:10.3390/w16172375.
- Zhu, X., *et al.*, 2024. Extreme precipitation accelerates nitrate leaching in the intensive agricultural region with thick unsaturated zones. *Science of the Total Environment*, 918, 170789.
- Zounemat-Kermani, M., *et al.*, 2021. Ensemble machine learning paradigms in hydrology: a review. *Journal of Hydrology*, 598, 126266.