Explainable Spatial Modeling of Groundwater Nitrate Concentrations in the Netherlands

Iulia Capralova[1], Juan Cardenas-Cartagena[1,*]

[1]University of Groningen, Faculty of Science and Engineering, Nijenborgh 9, 9747 AG Groningen, The Netherlands

Abstract

Nitrate leaching from soil is a major source of groundwater pollution that threatens both environmental and public health. In the Netherlands, intensive agricultural practices make the country vulnerable to nitrate contamination, requiring constant monitoring of groundwater quality. Since there are few measurement sites, there is a need for interpretable estimation methods via spatial modeling. This paper develops an explainable spatial regression model for estimating nitrate concentrations in groundwater in the Netherlands, using spatial and environmental factors. Three models were tested: Ridge Linear Regression, Random Forest, and Extreme Gradient Boosting (XGBoost). The ensemble of Random Forest and XGBoost explained 66% of the variance in nitrate levels, indicating competitive performance suitable for interpolation and map-based interpretation of spatial nitrate variability in the Netherlands. Key factors influencing nitrate leaching, including soil type, loam content in the soil, and groundwater depth, were identified using both model-specific and model-agnostic methods. Modeled spatial estimated maps show that between 2017 and 2023, nitrate concentrations decreased in the east and northeast, while they increased in the south and parts of the central Netherlands. This modeling framework can support targeted environmental policy decisions aimed at reducing groundwater pollution.

Keywords

Nitrate leaching, groundwater monitoring, geospatial analysis, spatial regression, explainable artificial intelligence.

1. Introduction

1.1. Motivation

Nitrate (NO_3^-) is a form of inorganic nitrogen that is found in soil and groundwater. It acts as a primary source of nitrogen for plants, supporting their growth and development [1]. Nitrate occurs naturally in groundwater through processes such as rock weathering and biological activity, resulting in a background level of approximately 2.5 mg/L [2]. However, in recent decades, nitrate levels in groundwater have increased worldwide due to human activities such as heavy use of nitrogen-based fertilizers and animal manure in areas with intensive farming. This causes nitrate to leak from the soil into the groundwater, leading to pollution of water bodies. Today, nitrate leaching is a serious environmental problem in many farming regions across Europe and beyond [3].

The imbalance of nitrate concentration in groundwater has consequences for both nature and human health. When nitrate enters surface waters such as lakes and rivers, it causes eutrophication (i.e., rapid growth of algae) which reduces oxygen levels and disrupts aquatic life [4]. Over time, this contamination affects the quality of drinking water, causing health issues such as methemoglobinemia and increasing the risk of gastric cancer [5, 6, 7, 8].

To protect water quality, the European Union (EU) created the Nitrates Directive (91/676/EEC), which sets a safe limit of 50 mg/L of nitrate concentration in groundwater [9]. It requires each member country to monitor nitrate levels, identify vulnerable areas, and take steps to reduce pollution. While this program has helped to improve the overall situation in many EU countries, in some local areas, nitrate contamination still remains an issue. The Netherlands, as one of the leading agricultural producers in the world, regularly monitors and reports nitrate levels through the Dutch National Minerals Policy

Workshop on AI-driven Data Engineering and Reusability for Earth and Space Sciences (DARES'25), co-located with the 28th European Conference on Artificial Intelligence (ECAI 2025), Bologna, Italy, October 25, 2025

*Corresponding author.

i.capralova@student.rug.nl (I. Capralova); j.d.cardenas.cartagena@rug.nl (J. Cardenas-Cartagena)

0009-0002-4121-0103 (I. Capralova); 0000-0001-9718-6929 (J. Cardenas-Cartagena)

Monitoring Program (LMM). Between 1992 and 2019, a 50% reduction in hotspot areas was achieved [10]. Despite this progress, challenges remain. For example, in sandy agricultural regions in the South of the Netherlands, shallow groundwater (5 - 15 meters) still exceeds the EU nitrate limit [11].

These localized spikes in nitrate levels indicate that policies should be tailored to local conditions rather than relying solely on national targets. To achieve this, authorities require an understanding of nitrate distribution across the country. However, monitoring nitrate levels across the country at high resolution and frequency remains a challenge. Taking nitrate samples is both expensive and time-consuming. Thus, it is measured once or twice per year at limited locations. This leaves gaps in our understanding and makes it difficult to identify high-risk areas.

This problem motivates the development and application of data-driven spatial regression models to explain how both natural conditions and human activities contribute to variations in these patterns across regions. These models can help identify polluted areas where measurements are unavailable, support groundwater protection plans, and inform more sustainable farming practices tailored to the needs of specific regions.

Furthermore, this study aligns with the United Nations Sustainable Development Goal 6, which aims to protect and use water resources responsibly. A key target under this goal is to improve water quality by reducing pollution, minimizing the release of hazardous chemicals, and halving the proportion of untreated wastewater [12]

1.2. State-of-the-Art

Nitrate leaching has been spatially estimated using geostatistical methods that use spatial and temporal autocorrelation, relying on Tobler's First Law of Geography, *"everything is related to everything else, but near things are more related than distant things"* [13]. Methods such as co-kriging and disjunctive kriging are based on this principle, as they estimate values at unsampled locations by leveraging the similarity between nearby observations. Incorporating temporal information has been shown to reduce estimation uncertainty further and improve spatial risk assessments [14, 15]. However, these methods require a minimum number of observations to compute reliable variograms[1], which limits their applicability in regions with sparse monitoring data. Moreover, focusing primarily on geographic proximity can result in omitting important environmental interactions and biogeochemical processes in the nitrogen cycle (see Section 2).

In recent years, machine learning (ML) has become a preferred method for environmental and soil studies. It can handle large datasets and capture complex patterns found in groundwater quality data. Among ML techniques, tree-based models are commonly used for spatial nitrate estimation due to their inherent model-specific interpretability and relatively low computational cost [16].

Paper written by Mahlknecht et al. [17] applied Extreme Gradient Boosting (XGBoost) to estimate nitrate concentrations across Mexico, achieving an accuracy of 0.80 despite limited monitoring data. They identified rainfall, elevation, and slope as key predictors and produced spatial risk maps that exposed major nitrate hotspots. Similarly, Covatti et al. [18] used Random Forest (RF) to map nitrate levels in Swiss groundwater, integrating SHAP values to interpret feature importance. The model explained 58% of the nitrate variance. The model revealed that factors like seasonal precipitation and soil organic carbon emerged as strong predictors. In the United States, Ransom et al. [19] developed a 3D XGBoost model to estimate nitrate across two drinking water zones, achieving an R^2 of 0.49 on hold-out data. Their analysis showed that well depth, soil and climate characteristics, and the absence of developed land were key factors influencing nitrate levels at the national scale. These studies demonstrate the ability of tree-based models to capture complex nitrate dynamics in the environment.

In the Dutch context, Spijker et al. [20] have developed an RF model to map nitrate leaching from agricultural soils using environmental predictors. The model explained 58% of the variance and highlighted land use and elevation as being key variables. However, the study was limited by its narrow temporal scope (restricted to 2017), reliance on a model-specific interpretability method only, and

[1] A tool in geostatistics used to analyze the spatial continuity of data by quantifying the degree of dissimilarity between data points as a function of distance.

the absence of key factors influencing the nitrogen cycle, such as population pressure, precipitation, temperature, and detailed soil properties, including chemical composition. These limitations are addressed in our study through broader environmental inputs, multi-year analysis, and model-agnostic interpretability methods.

1.3. Contributions

Building on the work of [20], our study aims to advance nitrate leaching estimation in the Netherlands by (1) increasing the temporal scope of both model training and estimation, (2) incorporating a broader range of environmental and agricultural predictors based on the nitrogen cycle, and (3) applying model-agnostic and model-specific interpretability techniques. Therefore, our research has the following objective: *The development of an explainable spatial regression model for estimating nitrate concentrations in groundwater in the Netherlands, using spatial and environmental data from agricultural soils.*

2. A Brief Introduction to the Nitrogen Cycle

The nitrogen cycle describes the natural processes through which nitrogen moves between the atmosphere, soil, water, and living organisms, which is summarized in Figure 1 [21]. Its understanding helps in feature selection and data preprocessing steps.

According to Fowler et al. [22], the nitrogen cycle begins when nitrogen enters the soil through atmospheric fixation, where lightning converts N_2 into reactive nitrogen compounds that are deposited with rainfall. Moreover, in agricultural areas, natural inputs are enhanced by the application of synthetic fertilizers, as well as organic sources such as livestock manure and urine. Additionally, urban areas primarily contribute through wastewater discharge from sewage systems into soil and water bodies. The effects are seen in densely populated regions, where higher concentrations of people are associated with greater wastewater production, fertilizer use, and nitrogen emissions [23]. These emissions, particularly ammonia from agriculture and nitrogen oxides from traffic and industry, settle onto the land as nitrogen deposition in soil.

Once in the soil, nitrogen undergoes a series of microbial transformations that convert it into forms plants can absorb. These processes are shaped by environmental conditions: soil texture influences oxygen and water flow, temperature affects microbial activity, and moisture levels determine whether nitrification or denitrification is favored. Eventually, part of this nitrate is taken up by plants, with uptake varying depending on the type of vegetation and the depth of the roots. For instance, according to Nouri et al. [24], deep-rooted or fast-growing crops tend to absorb more nitrates.

When the amount of nitrate in the soil exceeds the plant's uptake capacity, nitrate leaches into groundwater. Since nitrate (NO_3^-) is highly soluble in water and poorly retained by soil particles due to its negative charge, it is easily washed out from the root zone under the influence of precipitation, and groundwater table [22, 25]. Moreover, soil texture determines how water moves through the soil. For example, sandy soils, due to large particles, let water with nitrate pass through quickly [26, 27]. Conversely, the fine texture and low permeability of clay soils restrict water flow, which limits nitrate leaching [28]. Additionally, temperature regulates the activity of bacteria involved in nitrate production, thereby influencing nitrification and denitrification rates [29]. Elevation further influences nitrogen dynamics by shaping drainage, runoff, and the accumulation of water and nutrients in low-lying areas [30].

Apart from leaching, nitrogen can also be removed from the soil system through denitrification, a microbial process that transforms nitrate into gaseous nitrogen compounds (N_2, N_2O) released into the atmosphere. This process closes the nitrogen cycle by returning nitrogen to the atmosphere. The overall cycle offers insights into relevant data and features for developing explainable ML-based estimators for nitrate leaching.

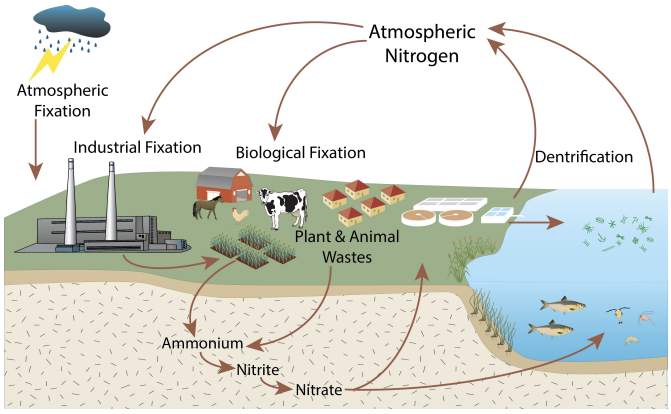


Figure 1: Nitrogen cycle –Simplified diagram [21].

3. Methods

This study models nitrate leaching in agricultural soils across the Netherlands by training regression models on georeferenced nitrate measurements and spatial-temporal features. Once trained, models are applied across a spatial grid, where each geographic location contains corresponding values for all predictors. The outcome is an estimated map, showing the spatial patterns and trends of nitrate leaching across the country. The following sections describe the data exploration analysis, the modeling framework, and the evaluation methods used to assess model performance.

3.1. Data Collection

The target variable is the nitrate concentration in groundwater. Measurements were obtained from well sensors logged in the DINO Loket platform [31]. The samples were collected according to the Dutch protocol NTA 8017 [32], which ensures standardization in sampling procedures and data quality. The sampling frequency for monitoring wells ranges from once to twice a year, which mostly falls in the winter months. Samples taken within city boundaries were excluded from the analysis.

Each well consists of multiple filters, which are screened intervals at different depths designed to isolate hydrogeological zones. For consistency and comparability across wells, we use data from Filter 1, the shallowest filter, which directly reflects nitrate leaching from the root zone of agricultural soils.

The predictor and target variables used in this study are summarized in Table 3 in Appendix A.2. Their selection was based on the Nitrogen Cycle (see Section 2). These datasets are derived from a range of sources and measurement techniques, which are described in Appendices A.4 and A.5. In total, the dataset contained 7620 samples and 18 features.

3.1.1. Study area

For this study, the focus is on the entire territory of the Netherlands, which covers an area of approximately 41,543 km^2 [33]. It lies within the Rhine–Meuse river basin, which explains relatively shallow groundwater levels [34]. The elevation ranges from 7 meters below sea level (in South Holland) to above 322 meters (in Limburg). The country has polders, which are areas of land taken from the sea or rivers and kept dry by dikes.

Unconsolidated Quaternary sediments of fluviatile, marine, or glacigenic origin dominate the subsurface of the Netherlands. The country is characterized by four main soil types: peat, clay, sand, and loess soils (see Figure 2a). Peat areas are mostly found in low-lying regions, clay soils dominate the coastal and riverine floodplains, sandy (50% of the area) soils are common in the eastern and southern parts of the country [35], and loess soil is found only in the South Netherlands (see Figure 2a).

Approximately 66% of the land is used for agriculture, while around 16% is covered by forests [36, 37]. The remaining area consists of urban settlements, water bodies, and infrastructure.

The Netherlands has a temperate maritime climate, with mild winters (3–6 °C), cool summers (17–20 °C), and a yearly average temperature of around 10 °C. Rainfall is evenly distributed throughout the year, averaging 800–900 mm annually, with most of the precipitation occurring in autumn and winter [38]. These conditions create a predisposition for nitrate leaching, especially after dry summers when nitrogen accumulates in the soil and is flushed out by winter rains [39].

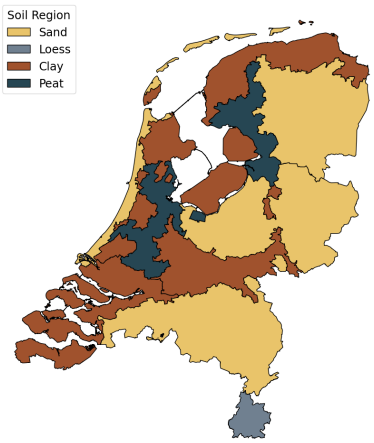
3.1.2. Data Inspection

To understand general trends in nitrate leaching and prepare for further analysis, we first look at the available data.

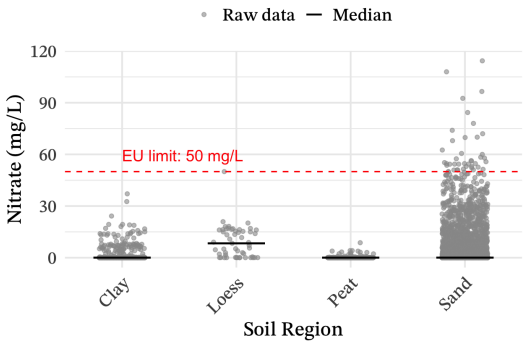
The monitoring network covers a total of 874 unique groundwater locations sampled between 2008 and 2023. Table 2 and Figure 9 in Appendix A provide an overview of the monitoring network. The table summarizes the number of unique sites sampled each year, with coverage peaking in 2015 (591 sites) and reaching its lowest in 2023 (274 sites), while the figure shows the combined spatial distribution of all monitoring sites across the country.

Figure 2b displays the distribution of nitrate concentrations across different soil regions: clay, loess, peat, and sand. In clay and peat regions, nitrate levels are low, with median concentrations of 0.05 mg/L and maximum values well below the EU legal limit of 50 mg/L. On the other hand, anaerobic conditions in peat lands support the conversion of nitrate back into nitrogen gas through denitrification. The loess region shows moderate nitrate levels, with a median of 8.37 mg/L, and some values reach the EU limit. In contrast, sandy soils have the highest nitrate concentrations and variability, with a median value of 0.07 mg/L but maximum values exceeding 100 mg/L, more than twice the EU threshold.

Appendix A.6 describes the preprocessing data pipeline designed to prepare training and test datasets for model training and evaluation.



(a) Map of soil regions in the Netherlands.



(b) Distribution of nitrate concentrations across soil regions.

Figure 2: Soil regions in the Netherlands and corresponding nitrate concentration distributions.

3.2. Model development

3.2.1. Algorithm selection

The goal of this study is to develop an explainable model of nitrate leaching across the Netherlands. For this reason, models with clear interpretations were prioritized. Three different models were chosen:

- **Ridge Linear Regression**, used as the baseline model, assumes a linear relationship between predictor and target variables. The resulting coefficients indicate both the direction (positive or negative) and strength of each variable's relationship with nitrate concentrations.
- **Random Forest**, proposed by Breiman [40], is an ensemble model that constructs a number of decision trees, each trained on random subsets of data and features.

- **XGBoost**, introduced by Chen and Guestrin [41], is a gradient boosting model that builds decision trees sequentially, where each new tree aims to correct the errors made by previous trees.

To combine the strengths of the best-performing models, an **ensemble** is created by taking an inverse-variance weighted average of the models' estimations [42].

All models are implemented using the scikit-learn library [43]. For XGBoost, the scikit-learn-compatible XGBRegressor interface from the XGBoost package is used [41].

3.2.2. Performance evaluation

In order to gain different perspectives on the model performance, the following evaluation metrics were used:

- Coefficient of determination (R^2) measures the proportion of variance in the target variable that is explained by the model. An R^2 value closer to 1 indicates better estimation performance:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1}$$

- Root Mean Square Error (RMSE) gives higher weight to large errors by squaring the residuals before averaging and taking the square root:

$$\text{RMSE}(y, \hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{2}$$

- Mean Absolute Error (MAE) calculates the average absolute difference between estimated and actual values:

$$\text{MAE}(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3}$$

In all three equations (1), (2), (3), y_i is the true value for sample i, $\hat{y}_i$ is estimated value for sample i, and n is number of samples. Additionally, in (1) $\bar{y}$ is the mean of the true values.

3.2.3. Hyperparameter optimization

To control the complexity and regularization strength of the chosen models, a hyperparameter search is performed. Specifically, the *Randomized Search Cross-Validation* implementation from scikit-learn [43] is used to sample 60 unique hyperparameter combinations from the predefined search space. Each configuration was evaluated using a 5-fold time series cross-validation scheme (*Time-Series Split*) to preserve temporal structure in the data and avoid information leakage between training and validation folds. Model performance during hyperparameter tuning was evaluated using MAE in (3). And R^2, in (1), was used as the refitting criterion for the final model selection.

Appendix B gives an overview of the hyperparameter search space along with the chosen ones for Ridge Linear regression (see Table 4), Random Forest (see Table 5), and XGBoost (see Table 6).

3.2.4. Model training

Once the optimal configuration for each model was identified, the corresponding model was retrained on the whole training set and evaluated on the held-out test set. The objective of the training procedure was to minimize the MAE, in (3). During model selection, R^2, in (1), was used as the refitting metric to prioritize models that explain more variance in the data. At the same time, RMSE, in (2), was also monitored as a secondary evaluation criterion.

To understand how each model generalizes as the amount of training data increases, a learning curve analysis was implemented based on repeated time series cross-validation. Once the best hyperparameters

were found for each model, the best-performing estimator was cloned and retrained across multiple increasing subsets of the training data. Specifically, for each training subset fraction, the corresponding data was split using a 10-fold *Time-Series Split*, and the model was fitted in each fold. Validation and training scores were computed using MAE and were stored across the folds to assess the learning pattern.

Resultant learning curves for Random Forest and XGBoost can be found in Appendix B on Figures 14a, and 14b respectively.

4. Results

4.1. Model results and estimated maps

The performance of each model was evaluated on the held-out test set (882 samples), using the three metrics described in Section 3.2.2.

As shown in Table 1, the Ridge Linear Regression model achieved the lowest performance, with an R^2 of 0.23, an MAE of 3.79 mg/L, and an RMSE of 7.58 mg/L. It was outperformed by both Random Forest and XGBoost with R^2 values of 0.649 and 0.638, MAEs of 2.10 mg/L and 2.04 mg/L, and RMSEs of 5.13 mg/L and 5.20 mg/L, respectively. Therefore, Random Forest and XGBoost were included in the Ensemble model, with weights of 0.52 and 0.48, respectively. Resultant ensemble achieved the best overall performance, with an R^2 of 0.66, an MAE of 2.02 mg/L, and an RMSE of 5.05 mg/L.

Table 1
Model performance comparison based on R^2, MAE, and RMSE metrics.

Model	R^2	MAE	RMSE
Ridge Linear Regression	0.23	3.79	7.58
Random Forest	0.649	2.10	5.13
XGBoost	0.638	2.04	5.20
Ensemble	**0.66**	**2.02**	**5.05**

To further assess the quality of each model, Figure 3 presents scatter plots comparing the estimated versus actual nitrate concentrations, and Figure 4 demonstrates the corresponding residuals.

In Ridge Linear Regression (see Figures 3a and 4a), estimations align with the diagonal for values below 10 mg/L. For higher concentrations (>15 mg/L), estimated values cluster below the diagonal, showing underestimation. Residuals increase with expected values, reaching up to 50 mg/L, and show a downward trend across the x-axis.

The estimation patterns of Random Forest (see Figures 3b and 4b) and XGBoost (see Figures 3c and 4c) are similar with minor differences. XGBoost estimates high nitrate concentrations (>40 mg/L) more accurately, while Random Forest performs better for mid-range values (20-30 mg/L). Residuals in XGBoost, in the range of 20-30 mg/L, are less clustered and more dispersed compared to those in Random Forest. Residual plots for both models show that residuals range approximately from -30 to 40 on the y-axis.

In the Ensemble model (see Figures 3d and 4d), the estimation and residual patterns are similar to those of Random Forest and XGBoost. For several points in the 10-30 mg/L and >40 mg/L ranges, residuals are smaller compared to those in the individual models. Some over- and underestimations present in RF and XGBoost are reduced.

Although the performance gains are small, the ensemble provides a more stable estimation. Thus, it is used for nitrate spatial modeling. Its performance on the test set of the year 2021 can be seen in Figure 5. The Ensemble model captures the spatial pattern of nitrate concentrations across the Netherlands,

specifically lower levels in the north and coastal areas, and high concentrations in the southern region of North Brabant (highlighted in the red square).

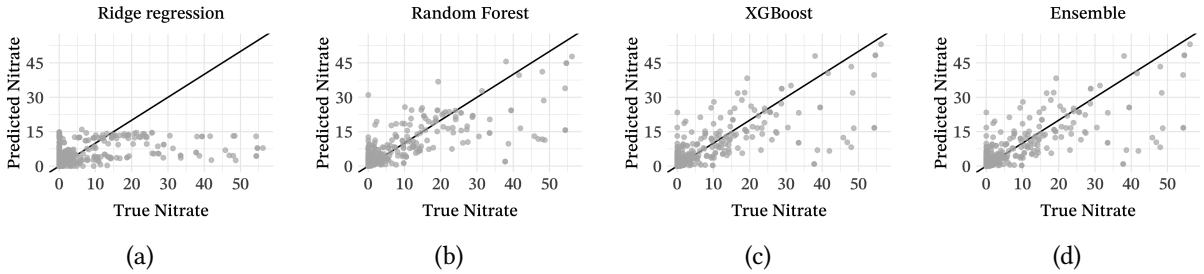


Figure 3: Estimated versus observed nitrate concentrations for (a) Ridge Linear Regression, (b) Random Forest, (c) XGBoost, and (d) the Ensemble model on the test set. The diagonal 1:1 line indicates perfect estimations.

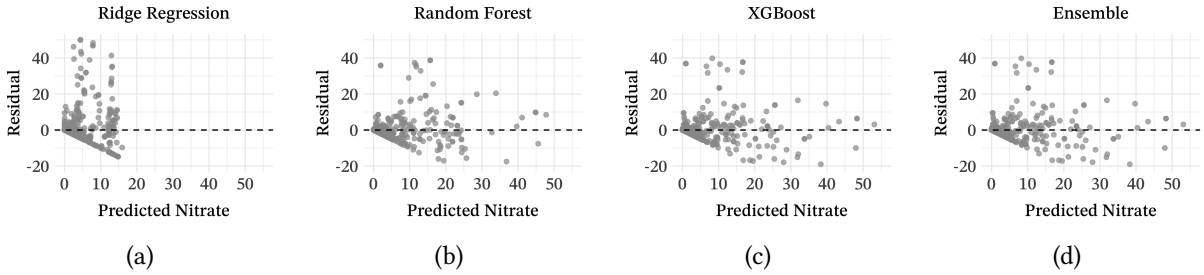


Figure 4: Residual plots for (a) Ridge Linear Regression, (b) Random Forest, (c) XGBoost, and (d) the Ensemble model on the test set. The dashed horizontal line indicates zero error.

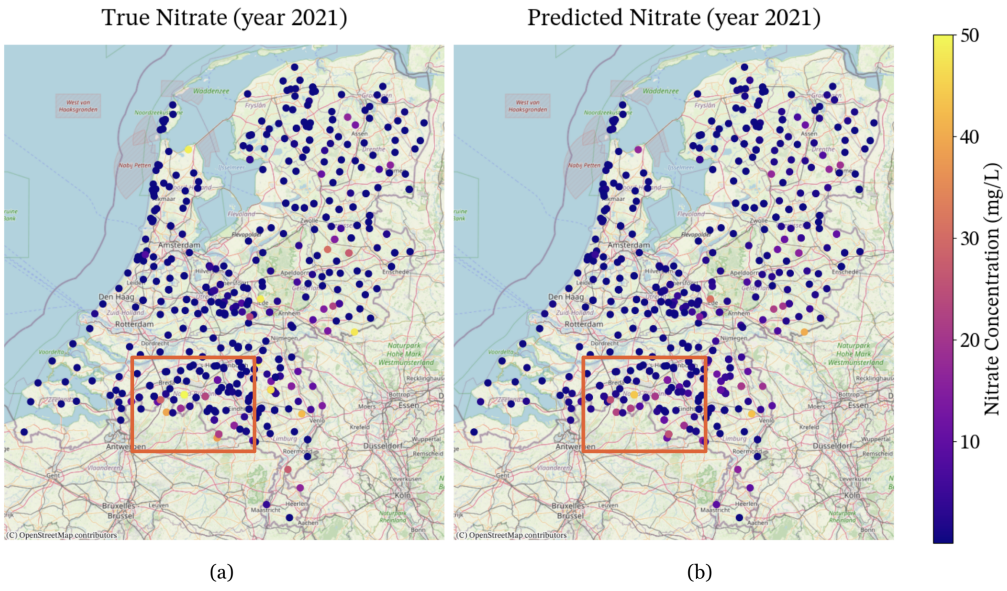


Figure 5: Observed (a) and estimated (b) nitrate concentrations in groundwater for the 2021 test set, using the Ensemble model. The red box highlights a hotspot in southern Noord-Brabant, which is captured by the model.

4.2. Spatial modeling

Ensemble, as the best performing model, was used with the predictor variables to generate nitrate maps at a resolution of 500 by 500 meters for the years 2010, 2017, 2021, and 2023 (see Figure 8).

Most of the Netherlands shows low nitrate concentrations, with values below 5 mg/L. The spatial pattern of increased nitrogen levels in region A (Noord-Brabant and Limburg) and region B (east

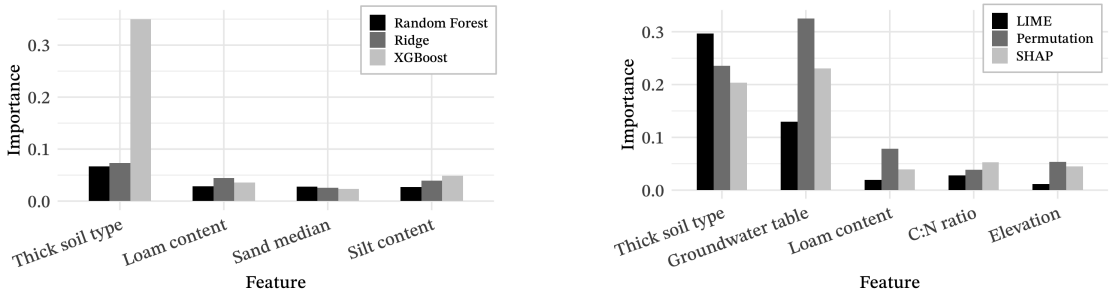
of Gelderland and Overijssel) remains stable through four maps. However, some changes in nitrate distribution and intensity are still present over time.

Figure 7 shows the difference in estimated nitrate concentrations for the years 2017, 2021, and 2023, each compared to 2010. In 2017 (see Figure 7a), nitrate concentrations increased in areas of region A (Noord-Brabant) and region B (Overijssel and Drenthe). In these areas, nitrate values are up to 10 mg/L higher than in 2010.

In 2021 (see Figure 7b), regions A and B still show an increase in nitrate levels. However, the pattern is different compared with 2017 (see Figure 7a). In region A, the increased nitrate remains, but in region B there is a significant decrease.

The pattern in both regions from 2021 is further developed in 2023 (see Figure 7c), with a slight decrease in nitrate levels in Region A and an even greater decline in Region B. However, a new hotspot with elevated nitrate concentrations emerged in the northern part of Drenthe.

4.3. Importance and influence of predictor variables



(a) Model-specific feature importance.

(b) Model-agnostic feature importance.

Figure 6: Comparison of feature importance results from (a) model-specific and (b) model-agnostic interpretability techniques.

To understand which predictors influence the nitrate leaching, feature importance was evaluated using model-specific and model-agnostic interpretability methods.

Figure 6a presents model-specific measures of feature relevance: Gini importance for both Random Forest and XGBoost, and feature coefficients for Ridge Linear Regression. Scores from each model were normalized to sum to 1, and only features shared across all three models are shown. Among them, "Thick soil type" is highlighted as the most important predictor. The models also agreed on the importance of "Loam content", "Sand median", and "Silt content".

Figure 6b presents the model-agnostic interpretability results based on LIME, Permutation Importance, and SHAP values. Across all three, "Thick soil type" and "Groundwater table" are found to be the most important features, although the importance values differ. For example, LIME gives "Thick soil type" the highest score (around 0.3), while SHAP and Permutation give it slightly lower scores. Moreover, all three methods further agreed on the importance of "Loam content", "C:N ratio", and "Elevation", however, with lower importance scores.

Taken together, both model-specific and model-agnostic approaches identified "Thick soil type" and "Loam content" as important predictors of nitrate leaching. Moreover, the strength of their importance is consistent across methods, with "Thick soil type" receiving the highest scores, and "Loam content" receiving the lowest ones.

5. Discussion

5.1. Interpretation of the results

The main goal of this study was to develop an explainable spatial regression model for estimating nitrate concentrations in groundwater in the Netherlands, using spatial and environmental data from soils.

Among all tested models, the Ensemble demonstrated the best performance, explaining 66% of the variance in nitrate levels with an RMSE of 5.05mg/L, which is considered strong for large-scale applications. It outperforms comparable studies such as Ransom et al. [19] in the United States (R^2 = 0.49), Covatti et al. [18] in Switzerland (R^2 = 0.58, RMSE = 7.8 mg/L), and Spijker et al. [20] in the Netherlands (R^2 = 0.58). As the main model framework was adapted from the approach by Spijker et al. [20], our results imply that increasing the training dataset and introducing new predictor variables, such as detailed soil properties, allowed the models to capture complex interactions more accurately.

From a practical perspective, this level of estimation performance means that the Ensemble model is suitable for spatial screening and groundwater risk assessment but should not be used as a precise indicator, as it consistently underestimates high nitrate concentrations (>30 mg/L). This is a common limitation of regression models [44]. Therefore, policy decisions near regulatory thresholds should not rely solely on model outputs but use them to guide targeted monitoring efforts.

5.2. Nitrate trend through years

Analyzing the spatial nitrate concentration distribution through the years based on Figure 8, it is clear that the lowest levels are found in peat, loess, and clay areas of the west and north of the country. However, as mentioned in Section 4.2, region A (North Brabant and Limburg) and region B (east of Gelderland and Overijssel) consistently show some of the highest modeled nitrate concentrations in the Netherlands, reaching 50 mg/L. This trend can be explained by the presence of sandy soils, which are known for their low capacity to retain water and nutrients [27]. These estimations coincide with [20] nitrate estimations and online reports from the LMM [45].

Based on the difference map of 2017 (see Figure 7a) the situation of high nitrate increased in Regions A and B. This can be explained by the increase in nitrogen surplus after 2015 [46] mainly due to higher emissions and greater use of artificial fertilizers, despite more nitrogen being removed with the crops.

In the following years, difference maps in 2021 in Figure 7b and 2023 in Figure 7c show that the increased nitrate concentrations persist in region A, while in region B the levels mostly return to the level of year 2010. This can be explained by the differences in sandy soils and land use practices. In Sand South, there is a predominance of dry, well-drained sandy soils (16% dry soils) and the relatively low proportion of reclaimed peat soils. Together they lead to faster nitrate leaching compared to wetter regions [47]. In addition, the agricultural structure differs across the sand subregions: Sand South has a lower share of grassland (dairy farms account for 46% of the area), which is associated with higher nitrogen leaching than grassland-dominated areas [48].

The decline in nitrate concentrations in Sand North and Sand Central after 2017 can be attributed to both environmental and agricultural factors. In Sand North, dairy farming accounts for 49%, while in Sand Central, dairy farming accounts for 67% [48, 45]. Both subregions have a relatively high share of grassland on dairy farms, which retains more nitrogen and results in lower nitrate leaching compared to arable land. These agricultural patterns, combined with a higher proportion of reclaimed peat soils in Sand North (49%), promote denitrification. This helps explain why nitrate concentrations peaked in 2017 but subsequently decreased, with Sand North even dropping below the EU threshold of 50 mg/L in recent years [49, 47].

The sharp nitrate increase observed in northern Drenthe in the difference map of 2023 (see Figure 7c) occurred in an area covered by coniferous and deciduous forests. This finding is unexpected, as forests absorb nitrate in large quantities, resulting in low nitrate leaching to groundwater [50]. This anomaly may be explained by the limitations of this study in Section 5.4.

5.3. Analysis of factors

Both model-agnostic and model-specific feature importance analyses identified the main soil classification, “Thick earth soil,” as the most important predictor of nitrate leaching in the Netherlands. Formed over centuries of deep fertilization and tillage, these soils develop a sandy top layer called the screed [51], which makes it easier for water and nutrients to move downward (see Section 2). This increases

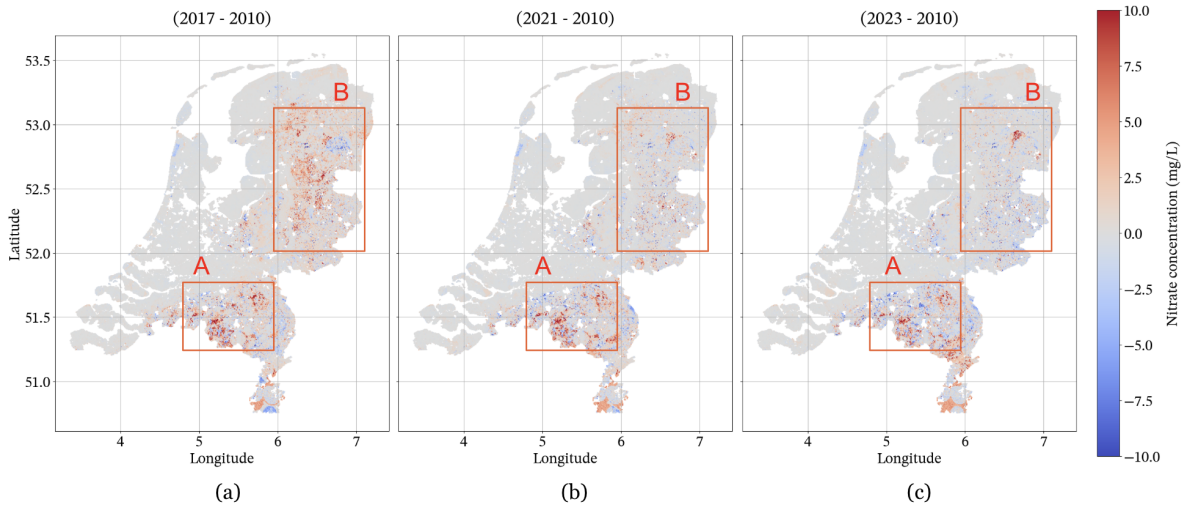


Figure 7: Estimated changes in nitrate concentrations between 2010 and later years based on ensemble model outputs: (a) 2017–2010, (b) 2021–2010, and (c) 2023–2010. The increased nitrate is in region A (Noord-Brabant) and region B (Drenthe and Overijssel).

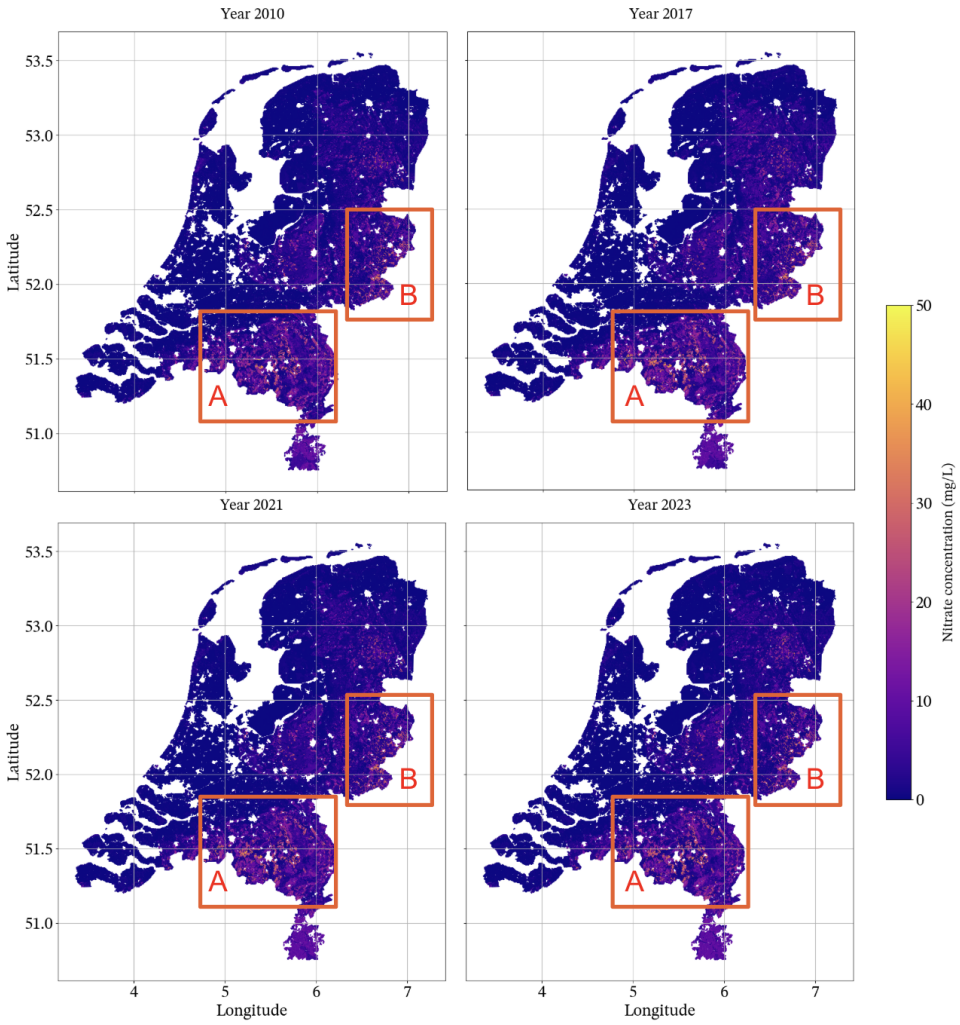


Figure 8: Estimated nitrate concentrations in groundwater across the Netherlands at a 500 × 500 m resolution for the years 2010, 2017, 2021, and 2023, based on the Ensemble model.

leaching risk, especially with heavy fertilizer use and irrigation. Although thick soils can store more nitrate, their structure also allows it to reach groundwater quickly. These soils are mainly found in the higher cover sand regions of Brabant, Gelderland, and Overijssel, matching the areas with the highest estimated nitrate levels (see Figure 8).

Loam content also emerged as a key predictor. Loam, a balanced mix of sand, silt, and clay, holds water and nutrients well while still draining, which slows nitrate leaching. In this study, lower loam content was linked to higher nitrate levels, suggesting that sandier, more permeable soils raise leaching risk, consistent with previous findings [27].

Groundwater depth was also identified as important by the model-agnostic method. SHAP values show that greater groundwater depth is associated with higher nitrate concentrations, which may seem counterintuitive, since deeper groundwater could allow more nitrate to be filtered through the soil [25]. However, in the Netherlands, deeper groundwater tables are typically found in elevated sandy regions with intensive agriculture and high nitrate inputs, as also noted by Spijker et al. [20]. Thus, groundwater depth here is likely an indicator of broader landscape characteristics.

5.4. Limitations

The spatial mismatch between nitrate and groundwater measurements (see Appendix A.6) may introduce uncertainty, as reflected by the moderate correlation between these variables (maximum correlation coefficient of $r = 0.6$). This suggests that spatial alignment issues could weaken the precision of the model's relationships. Also, missing land use data for some years was filled using maps from nearby years, which may have hidden short-term changes affecting nitrate leaching. And, while the ensemble model explained 66% of the variance in nitrate concentrations, a substantial proportion of variability remains unexplained. This residual variance likely arises from missing explanatory variables, such as detailed fertilizer application data. While the EU's Farm Accountancy Data Network (FADN) does collect farm-level information on fertilizer usage, only aggregated data is publicly available [52]. This lack of high-resolution fertilizer data restricts the ability to capture field-level variations in nitrate input. These factors highlight the need for more detailed spatial and temporal data, and the development of more comprehensive models to further improve the estimation of nitrate leaching in Dutch groundwater.

6. Conclusions

This project developed an explainable machine learning models for estimating nitrate concentrations in Dutch groundwater, utilizing a combination of spatial and environmental data. The best performing model is the Ensemble model that could explain 66% of the variance in nitrate concentrations. By integrating both model-specific and model-agnostic interpretability methods, the study identified thick type of soils, loam content, and groundwater depth as the most important predictors of nitrate leaching risk. The results highlight the vulnerability of elevated sandy regions. Moreover, the results show that overall nitrate concentrations remain low across most of the country, with high values only in a few regions, suggesting that the EU regulatory limit has been effective in protecting groundwater quality.

There are two main directions for future research. First, to improve predictive performance and better explain nitrate variability, future models should integrate additional predictors related to local agricultural practices and fertilizer application. Second, in this study the validation methods do not account for spatial patterns. While a time-based split was applied to prevent temporal leakage, the training and test samples remain geographically close to one another. This spatial proximity may introduce correlation, which may result in performance estimates that do not accurately reflect the models' ability to generalize to new locations. Therefore, a spatial k-fold cross-validation proposed by Pohjankukka et al. [53] should be implemented.

Building on the findings of this project, future development should prioritize decision-support tools based on these machine learning models to guide local nitrate management and policy decisions. Such a tool could help policymakers and citizens to identify areas most at risk and tailor interventions more precisely.

Reproducibility

For information about the datasets utilized, refer to Section 3.1 and Table 3 in Appendix A.2. The code and experiments of this project can be found in the repository: github.com/IuliaCapralova/Nitrate-Spatial-Modeling-NL.

Declaration on Generative AI

While writing this paper, the authors used ChatGPT (OpenAI) and Grammarly for formatting assistance and to improve the writing style. All suggestions from these tools were reviewed and edited as needed by the authors, who take full responsibility for the final content.

Acknowledgments

The authors thank Dr. P. Pradhan and Dr. J. Spijker for their discussions on environmental challenges in the Netherlands, Dr. H. de Weerd for his feedback on the analysis presented in this work, and the University of Groningen for providing access to the Hábrók High Performance Computing cluster.

References

[1] C. Muratore, L. Espen, B. Prinsi, Nitrogen uptake in plants: the plasma membrane root transport systems from a physiological and proteomic perspective, Plants 10 (2021) 681.

[2] N. M. Dubrovsky, K. R. Burow, G. M. Clark, J. M. Gronberg, P. A. Hamilton, K. J. Hitt, D. K. Mueller, M. D. Munn, B. T. Nolan, L. J. Puckett, et al., The quality of our nation's waters: Nutrients in the nation's streams and groundwater, 1992-2004, Technical Report, US Geological Survey, 2010.

[3] E. Craswell, Fertilizers and nitrate pollution of surface and ground water: an increasingly pervasive global problem, SN Applied Sciences 3 (2021) 518.

[4] J. A. Camargo, Á. Alonso, Ecological and toxicological effects of inorganic nitrogen pollution in aquatic ecosystems: a global assessment, Environment international 32 (2006) 831–849.

[5] S. Suthar, P. Bishnoi, S. Singh, P. K. Mutiyar, A. K. Nema, N. S. Patil, Nitrate contamination in groundwater of some rural areas of rajasthan, india, Journal of hazardous materials 171 (2009) 189–199.

[6] L. Fewtrell, Drinking-water nitrate, methemoglobinemia, and global burden of disease: a discussion, Environmental health perspectives 112 (2004) 1371–1374.

[7] H.-F. Chiu, C.-H. Kuo, S.-S. Tsai, C.-C. Chen, D.-C. Wu, T.-N. Wu, C.-Y. Yang, Effect modification by drinking water hardness of the association between nitrate levels and gastric cancer: evidence from an ecological study, Journal of Toxicology and Environmental Health, Part A 75 (2012) 684–693.

[8] R. Picetti, M. Deeney, S. Pastorino, M. R. Miller, A. Shah, D. A. Leon, A. D. Dangour, R. Green, Nitrate and nitrite contamination in drinking water and cancer risk: A systematic review with meta-analysis, Environmental Research 210 (2022) 112988.

[9] C. Directive, et al., Concerning the protection of waters against pollution caused by nitrates from agricultural sources, Official Journal 375 (1991).

[10] J. Serra, C. Marques-dos Santos, J. Marinheiro, S. Cruz, M. Cameira, W. De Vries, T. Dalgaard, N. Hutchings, M. Graversgaard, F. Giannini-Kurina, et al., Assessing nitrate groundwater hotspots in europe reveals an inadequate designation of nitrate vulnerable zones, Chemosphere 355 (2024) 141830.

[11] J. Claessens, D. van Gils, T. Brussée, R. van Duijnen, M. Oosterwoud, A. Vrijhoef, A. Plette, M. Kotte, J. Rozemeijer, K. Ouwerkerk, et al., Agricultural practices and water quality in the netherlands: status (2020–2023) and trends (1992–2023): the 2024 nitrate report with the results of the monitoring of the effects of the eu nitrates directive action programmes (2024).

[12] United Nations, Goal 6: Ensure access to water and sanitation for all, 2015. URL: https://www.un.org/sustainabledevelopment/water-and-sanitation/.
[13] W. R. Tobler, A computer movie simulating urban growth in the detroit region, Economic geography 46 (1970) 234–240.
[14] V. D'Agostino, E. Greene, G. Passarella, M. Vurro, Spatial and temporal study of nitrate concentration in groundwater by means of coregionalization, Environmental geology 36 (1998) 285–295.
[15] G. Passarella, M. Vurro, V. D'agostino, G. Giuliano, M. J. Barcelona, A probabilistic methodology to assess the risk of groundwater quality degradation, Environmental Monitoring and Assessment 79 (2002) 57–74.
[16] B. T. Nolan, M. N. Fienen, D. L. Lorenz, A statistical learning framework for groundwater nitrate models of the central valley, california, usa, Journal of Hydrology 531 (2015) 902–911.
[17] J. Mahlknecht, J. A. Torres-Martínez, M. Kumar, A. Mora, D. Kaown, F. J. Loge, Nitrate prediction in groundwater of data scarce regions: The futuristic fresh-water management outlook, Science of the Total Environment 905 (2023) 166863.
[18] G. Covatti, K.-Y. Li, J. Podgorski, L. H. Winkel, M. Berg, Nitrate contamination in groundwater across switzerland: Spatial prediction and data-driven assessment of anthropogenic and environmental drivers, Science of the Total Environment 973 (2025) 179121.
[19] K. M. Ransom, B. T. Nolan, P. Stackelberg, K. Belitz, M. S. Fram, Machine learning predictions of nitrate in groundwater used for drinking supply in the conterminous united states, Science of the Total Environment 807 (2022) 151065.
[20] J. Spijker, D. Fraters, A. Vrijhoef, A machine learning based modelling framework to predict nitrate leaching from agricultural soils across the netherlands, Environmental Research Communications 3 (2021) 045002.
[21] C. Ward, The nitrogen cycle with haber-bosch process, https://ian.umces.edu/media-library/, 2013. Integration and Application Network, University of Maryland Center for Environmental Science. Licensed under CC BY-SA 4.0. Accessed: 2025-07-19.
[22] D. Fowler, M. Coyle, U. Skiba, M. A. Sutton, J. N. Cape, S. Reis, L. J. Sheppard, A. Jenkins, B. Grizzetti, J. N. Galloway, et al., The global nitrogen cycle in the twenty-first century, Philosophical Transactions of the Royal Society B: Biological Sciences 368 (2013) 20130164.
[23] B. T. Nolan, B. C. Ruddy, K. J. Hitt, D. R. Helsel, A national look at nitrate contamination of ground water, Water conditioning and purification 39 (1998) 76–79.
[24] A. Nouri, S. Lukas, S. Singh, S. Singh, S. Machado, When do cover crops reduce nitrate leaching? a global meta-analysis, Global Change Biology 28 (2022) 4736–4749.
[25] Q. Chen, A. Chen, J. Min, L. Li, W. Hu, C. Wang, B. Fu, S. Guo, D. Zhang, Shallow groundwater table fluctuations weaken nitrogen accumulation in the thin layer vadose zone of cropland around plateau lakes, southwest china, Science of The Total Environment 950 (2024) 175300.
[26] H. M. Van Es, J. M. Sogbedji, R. R. Schindelbeck, Effect of manure application timing, crop, and soil type on nitrate leaching, Journal of environmental quality 35 (2006) 670–679.
[27] H. Domnariu, C. Paltineanu, D. Marica, A.-R. LăcătuȘU, N. Rizea, R. Lazăr, G. A. Popa, A. Vrinceanu, C. Bălăceanu, Influence of soil-texture on nitrate leaching from small-scale lysimeters toward groundwater in various environments, Carpathian Journal of Earth and Environmental Sciences 15 (2020) 301–310.
[28] S. Wang, B. Wu, Y. Wang, X. Wang, Nitrate migration and transformation in low permeability sediments: laboratory experiments and modeling, Water 15 (2023) 2528.
[29] J. Lapierre, P. V. F. Machado, Z. Debruyn, S. E. Brown, S. Jordan, A. Berg, A. Biswas, H. A. Henry, C. Wagner-Riddle, Winter warming effects on soil nitrate leaching under cover crops: A field study using high-frequency weighing lysimeters, Frontiers in Environmental Science 10 (2022) 897221.
[30] S. J. Hall, C. G. Tenesaca, N. C. Lawrence, D. I. Green, M. J. Helmers, W. G. Crumpton, E. A. Heaton, A. VanLoocke, Poorly drained depressions can be hotspots of nutrient leaching from agricultural soils, Technical Report, Wiley Online Library, 2023.

[31] G. S. o. t. N. TNO, Dino loket: Data and information of the dutch subsurface, https://www.dinoloket.nl/, 2025. Accessed: 2025-07-19.
[32] NEN, NTA 8017:2016 – Sampling of groundwater for the purpose of monitoring groundwater quality, https://www.nen.nl, 2016. Accessed: 2025-07-20.
[33] Statistics Netherlands (CBS), Population density; postcode and address, 1 january 2023, https://opendata.cbs.nl/statline/#/CBS/en/dataset/70262ENG/table?dl=5EFAE, 2023. Accessed on 20-07-2025.
[34] E. H. Sutanudjaja, L. Van Beek, S. M. De Jong, F. C. van Geer, M. Bierkens, Large-scale groundwater modeling using global datasets: a test case for the rhine-meuse basin, Hydrology and Earth System Sciences 15 (2011) 2913–2935.
[35] Delta Programme, Elevated sandy soils, https://english.deltaprogramma.nl/areas/elevated-sandy-soils, 2024. Accessed: 2025-07-23.
[36] Eurostat, Land cover statistics, https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Land_cover_statistics, 2025. Accessed: 2025-07-23.
[37] European Commission, Cap strategic plan: Netherlands, https://agriculture.ec.europa.eu/cap-my-country/cap-strategic-plans/netherlands_en, 2025. Accessed: 2025-07-23.
[38] World Bank, Climate data - historical: Netherlands, https://climateknowledgeportal.worldbank.org/country/netherlands/climate-data-historical, 2024. Accessed on 12-08-2025.
[39] I. Raij-Hoffman, O. Dahan, H. E. Dahlke, T. Harter, I. Kisekka, Assessing nitrate leaching during drought and extreme precipitation: Exploring deep vadose-zone monitoring, groundwater observations, and field mass balance, Water Resources Research 60 (2024) e2024WR037973.
[40] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.
[41] T. Chen, C. Guestrin, Xgboost: A scalable tree boosting system, in: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785–794.
[42] J. Hartung, G. Knapp, B. K. Sinha, Statistical meta-analysis with applications, John Wiley & Sons, 2011.
[43] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, the Journal of machine Learning research 12 (2011) 2825–2830.
[44] G. Zhang, Y. Lu, Bias-corrected random forests in regression, Journal of Applied Statistics 39 (2012) 151–160.
[45] Rijksinstituut voor Volksgezondheid en Milieu (RIVM), Nitraatmetingen - lmm tabel 2022, https://lmm.rivm.nl/Tabel/2022/Nitraat, 2022. Accessed: 2025-07-22.
[46] B. Fraters, A. Hooijboer, A. Vrijhoef, A. Plette, N. Van Duijnhoven, J. Rozemeijer, M. Gosseling, C. Daatselaar, J. Roskam, H. Begeman, Agricultural practices and water quality in the netherlands; status (2016-2019) and trend (1992-2019) (2021).
[47] O. Schoumans, P. Groenendijk, L. Renaud, W. van Dijk, J. Schroder, A. van den Ham, A. Hooijboer, Verhoogde nitraatconcentraties in het zuidelijke zandgebied: analyse van de mogelijke oorzaken (2012).
[48] Alterra, De bodemkaart van nederland, schaal 1:50 000, http://www.bodemdata.nl/ (accessed July 22, 2025), 2006.
[49] RIVM, Rijksinstituut voor Volksgezondheid en Milieu (RIVM), Grootschalige Depositiekaarten Nederland – historische reeks (2005–2023), RIVM open data portal, 2024. URL: https://www.rivm.nl/gcn-gdn-kaarten/depositiekaarten/historische-reeksen, annual nitrogen and acid deposition maps at 1 km² resolution from 2005–2023; downloadable as zip archives with accompanying readme files.
[50] A. Rothe, Tree species management and nitrate contamination of groundwater: a central european perspective, in: Tree Species Effects on Soils: Implications for Global Change: Proceedings of the NATO Advanced Research Workshop on Trees and Soil Interactions, Implications to Global Climate Change August 2004 Krasnoyarsk, Russia, Springer, 2005, pp. 71–83.
[51] Stiboka / Bodemdata.nl, Dikke eerdgronden, https://legenda-bodemkaart.bodemdata.nl/bodemclassificatie/item/dikke-eerdgronden, 2025. Accessed: 2025-07-24.

[52] European Commission, DG-Agriculture and Rural Development, Fertiliser – agri-food data portal, https://agridata.ec.europa.eu/extensions/DataPortal/fertiliser.html, 2025. Accessed: 2025-07-24.
[53] J. Pohjankukka, T. Pahikkala, P. Nevalainen, J. Heikkonen, Estimating the prediction performance of spatial models via spatial k-fold cross validation, International Journal of Geographical Information Science 31 (2017) 2001–2019.
[54] WUR, Wageningen Environmental Research BRO Grondwaterspiegelmodel, 2024-01, https://www.dinoloket.nl/ondergrondkleding/kaart, 2024. Accessed on 29-06-2025.
[55] Rijkswaterstaat, Hoogtebestand ijssel 50 cm 2023 dtm, 2023.
[56] C. B. v. d. S. CBS, CBS Vierkantstatistieken 500 m, PDOK / INSPIRE Geoportal, 2025. URL: https://www.pdok.nl/introductie/-/article/cbs-vierkantstatistieken-500m, statistical data per 500 × 500 m grid cell for the Netherlands, peiljaar 1971–2024; extended June 4, 2025 with 2024 data.
[57] RIVM, Historische reeks stikstofdepositie (gpkg), versie 20241001, https://data.rivm.nl/data/stikstof/Stikstof%20Natuur/2024/, 2024. Accessed on 29-06-2025.
[58] KNMI, Royal netherlands meteorological institute data platform, https://dataplatform.knmi.nl/, 2024. Accessed on 29-06-2025.
[59] WUR, Bodemkaart van nederland v2024-01, https://www.broloket.nl/ondergrondmodellen, 2024. Accessed on 29-06-2025.
[60] WUR, Landelijk Grondgebruiksbestand Nederland 2021 (LGN2021): achtergronden, methodiek en validatie, Rapport / Wageningen Environmental Research 3235, Wageningen Environmental Research, Wageningen, Netherlands, 2023. URL: https://doi.org/10.18174/585714. doi:10.18174/585714, cC BY-SA 4.0 licence; grid with 51 land use classes, 5 m resolution.
[61] RIVM, Landelijk meetnet effecten mestbeleid (lmm) dataset, https://data.rivm.nl/data/lmm/, 2024. Accessed on 29-06-2025.
[62] KNMI, Uitleg over automatisch weerstations, https://www.knmi.nl/kennis-en-datacentrum/uitleg/automatische-weerstations, 2024. Accessed on 30/06/2024.
[63] V. L. Mulder, M. J. Hack-ten Broeke, M. van Doorn, K. Teuling, et al., BIS-4D: Maps of soil properties and their uncertainties at 25 m resolution in the Netherlands. Version 3, 2024. URL: https://doi.org/10.4121/0c934ac6-2e95-4422-8360-d3a802766c71.v3. doi:10.4121/0c934ac6-2e95-4422-8360-d3a802766c71.v3, dataset.
[64] NRCS, Soil health literature summary: Effects of conservation practices on soil properties in areas of cropland (2015).
[65] R. Jandl, E. Leitgeb, M. Englisch, Decadal changes of organic carbon, nitrogen, and acidity of austrian forest soils, Soil Systems 6 (2022) 28.
[66] V. P. Kaandorp, H. P. Broers, Y. Van Der Velde, J. Rozemeijer, P. G. De Louw, Time lags of nitrate, chloride, and tritium in streams assessed by dynamic groundwater flow tracking in a lowland landscape, Hydrology and Earth System Sciences 25 (2021) 3691–3711.
[67] D. McKay Fletcher, S. Ruiz, K. Williams, C. Petroselli, N. Walker, D. Chadwick, D. L. Jones, T. Roose, Projected increases in precipitation are expected to reduce nitrogen use efficiency and alter optimal fertilization timings in agriculture in the south east of england, ACS Es&t Engineering 2 (2022) 1414–1424.
[68] R. P. Ribeiro, N. Moniz, Imbalanced regression and extreme value prediction, Machine Learning 109 (2020) 1803–1835.

Appendices

A. Data insights

This appendix further investigates the dataset by analyzing correlations between target and predictor variables and examining time series and spatial covariates. It also includes an overview of data sources, availability, key summary statistics, and highlights outliers through visualization.

A.1. Spatial distribution of monitoring wells

This section focuses on an overview of the spatial distribution of monitoring locations. Figure 9 shows the combined distribution of all locations on the map, while Table 2 lists the number of unique monitoring locations per year. The wells are spread throughout the country, with higher densities in provinces such as Utrecht, North-Brabant, and North Holland.

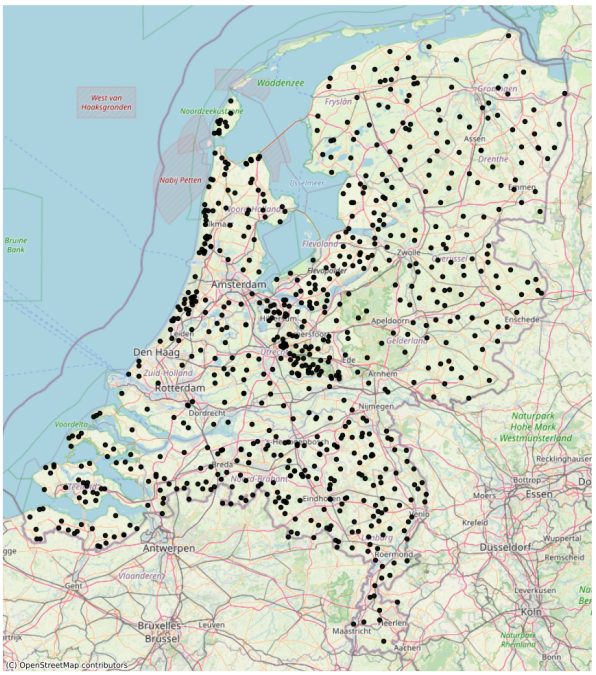


Figure 9: Monitoring well locations across all years (2008 - 2023) in the Netherlands.

Table 2
Number of unique monitoring locations per year across the territory of the Netherlands.

Year	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Locations	321	452	464	427	498	393	359	591	349	404	570	425	264	438	283	274

A.2. Overview of Variables Used in Modeling

Table 3
List of predictor variables and the target variable (nitrate), with units and data sources.

Variable	Unit	Reference
Nitrate (NO_3^-)	mg/L	WUR [54]
Elevation	m	Rijkswaterstaat [55]
Population	persons/km^2	CBS [56]
Groundwater table	m	WUR [54]
Nitrogen deposition	kg/ha/year	RIVM [57]
Precipitation	0.1 mm	KNMI [58]
Temperature	0.1 °C	KNMI [58]
Organic matter	%	WUR [59]
Land use	Category	WUR [60]
Soil region	Category	RIVM [61]
Soil type	Category	WUR [59]
C:N ratio	mol ratio	WUR [59]
Soil calcic content	%	WUR [59]
Fe-idth	%	WUR [59]
Soil loam content	%	WUR [59]
Sand median diameter	μm	WUR [59]
Sand silt content	%	WUR [59]
Soil acidity	pH	WUR [59]
Bulk density	g/cm^3	WUR [59]

A.3. Correlation map

To explore the relationships between nitrate concentrations and the explanatory variables presented in Table 3, a Pearson correlation analysis was conducted. Since Pearson correlation is only meaningful for continuous variables, the analysis was limited to these features. Figure 10 shows the resulting correlation matrix based on the absolute values of pairwise correlations, allowing comparison of both positive and negative relationships in terms of strength.

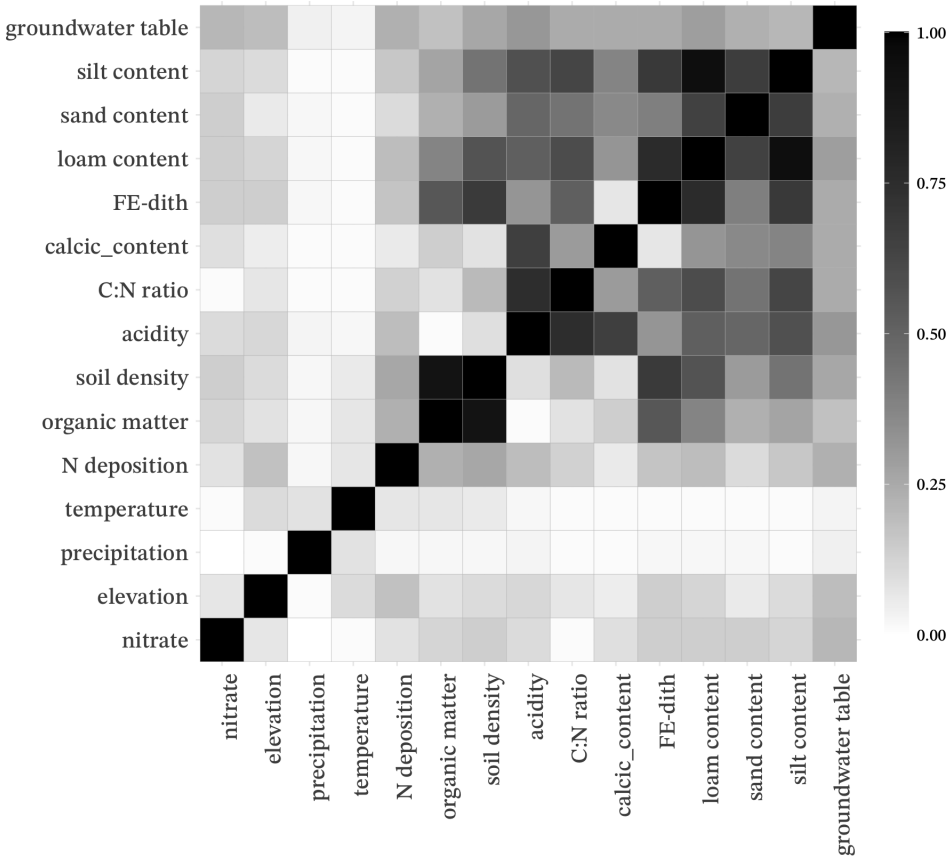


Figure 10: Pearson correlation matrix of nitrate and predictors. The strongest correlations are observed between loam content and silt content, sand content and silt content, soil acidity with C:N ratio, and organic matter with soil density. Nitrate shows weak correlations with most predictors, suggesting a complex relationship with multiple interacting factors.

A.4. Exploration of timeseries covariates

Groundwater table Groundwater table was included as a key explanatory variable as its level may accelerate nitrate washing out from the soil [25]. The dataset consisted of groundwater depth measurements collected from monitoring wells between 2008 and 2023. The measurements were obtained from a network of permanent monitoring wells equipped with automated sensors (pressure transducers) that recorded water levels at sub-daily to daily intervals. Although only a few wells had uninterrupted data across all years, the analysis included any well with a continuous ("clean") segment of measurements within this period, provided the segment exceeded 60 days. Given the high density of wells available, only the longest segment per well was used, and the rest were discarded to avoid redundancy. The raw data from each well often featured variable sampling frequencies, sometimes with intervals shorter than a day and with occasional missing periods. To standardize the data and address missing values, the depth measurements were first resampled to a uniform daily frequency using mean aggregation. Short gaps of up to five days were then filled by linear interpolation, making the time series consistent. In case this gap is exceeded, rows were considered as missing. The distribution of the final groundwater table dataset is shown in Figure 12a.

One challenge in this study was that the locations where groundwater depth was measured did not always match the nitrate measurement sites. To solve this issue, each nitrate measurement was linked to the closest groundwater well within 5 km. This radius was chosen based on the correlation between

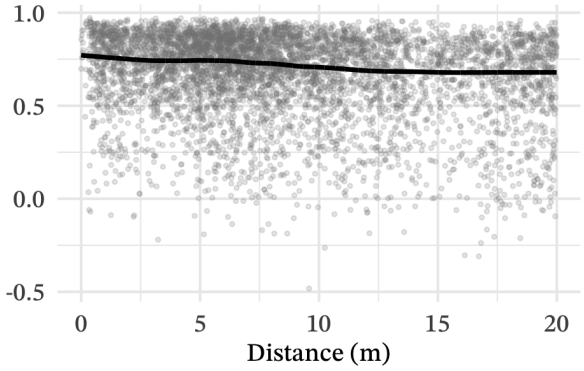


Figure 11: Scatterplot showing the correlation between groundwater depth and distance from the nearest nitrate measurement location, with a fitted trend line.

the distance between wells and groundwater table (see Figure 11). While this method made it possible to use more data, it also added some uncertainty. The correlation between groundwater depth and the distance reaches a maximum value of 0.61, highlighting this limitation.

Temperature and Precipitation Precipitation and temperature were included as explanatory variables to capture the influence of weather conditions on soil processes and nitrate mobility. Daily meteorological data were obtained from the Royal Netherlands Meteorological Institute [62], which operates a network of automatic weather stations (AWSs) that provide synoptic, high-quality measurements validated according to NEN-EN-ISO/IEC 17025 standards. The original data were recorded at an hourly frequency and were resampled to daily averages prior to further processing. Data were available for the period 2008–2023 and were sourced from 18 stations distributed across the area of interest. Each nitrate sampling location to the nearest weather station within a 20 km radius. The distribution of both precipitation and temperature across all stations is shown in Figure 12b and 12c, respectively.

For each unique sampling point, aggregated precipitation and temperature values were computed based on the nearest available station. Next, both temperature and precipitation variables were aggregated using the autocorrelation function (ACF) analysis as described in Appendix A.6.

A.5. Exploration of spatial covariates

Land use The land use data were sourced from the LGN (Landelijk Grondgebruiksbestand Nederland) dataset [60], which provides nationwide, high-resolution maps derived from a combination of satellite imagery and aerial photography. For more recent years, Sentinel-2 satellite data and imagery from the National Satellite Data Portal were utilized, while the 2021 update also included detailed aerial photographs. The spatial resolution of the maps is 5 meters, enabling precise assignment of land use classes to each sampling location. The land use map was available for the following years: 2008, 2012, 2018, 2019, 2020, 2021, 2022, and 2023. For years, lacking a land use map, the closest available year was adopted:

1. 2008 data was used for 2009
2. 2012 data was used for 2010, 2011, 2013, and 2014
3. 2018 data was used for 2015, 2016, and 2017

Nitrogen (N) deposition Annual spatial maps of nitrogen deposition for the period 2008–2023 were obtained from the national Data on Deposition in the Netherlands (GDN), developed by the Dutch National Institute for Public Health and the Environment [57]. These maps provide estimates of both wet and dry deposition, which are major sources such as ammonia NH_3 and nitrogen oxides NOx. Deposition values are modeled using the OPS atmospheric transport model, which integrates official

emission inventories, land-use information, and year-specific meteorological data. For accuracy, modeled estimates are calibrated against ground-based measurements from the national monitoring network, employing a co-kriging approach that addresses spatial uncertainty and variations in measurement reliability. The resulting calibrated maps, with a spatial resolution of 250 x 250 meters, provide a spatial representation of nitrogen inputs, distribution of which is available in Figure 12d.

Population density Annual gridded population maps for the years 2008–2023 were sourced from Statistics Netherlands (CBS) [33], with a spatial resolution of 500 by 500 meters. The dataset is based on the Dutch Personal Records Database (BRP), which provides register-based demographic information and has replaced traditional census approaches. In accordance with privacy regulations, grid squares containing fewer than five residents are not published. For the purposes of this study, such missing values were imputed with zeros, reflecting uninhabited areas such as forests, water bodies, or agricultural fields. The distribution of population density across nitrate sampling locations is shown in Figure 12e.

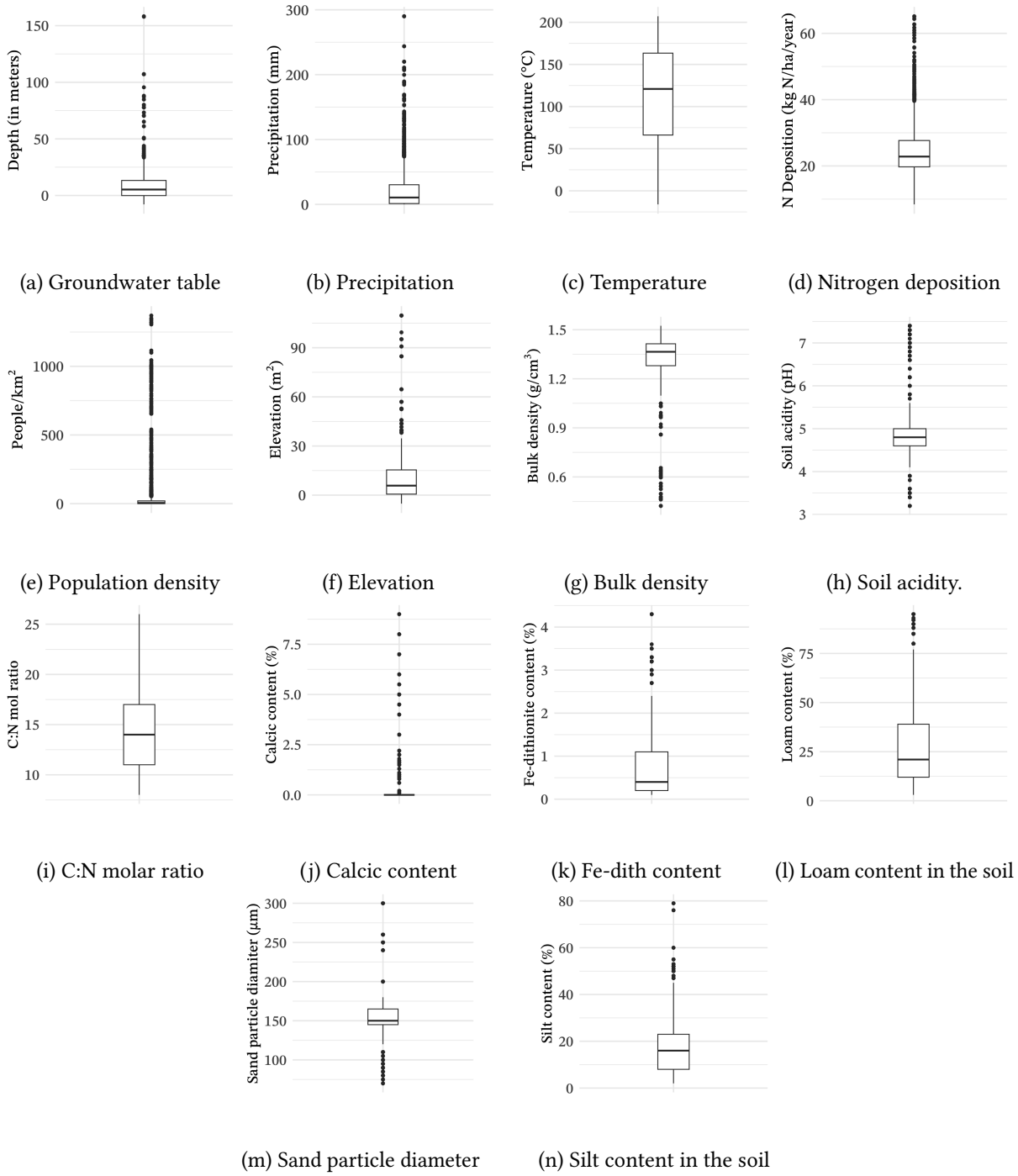
Elevation Elevation data were sourced from the Actueel Hoogtebestand Nederland (AHN) [55], a national digital elevation model produced using laser altimetry from aircraft. The accuracy of the map is provided by GPS corrections and merging overlapping data swaths. The resultant map has 0.5 by 0.5 meter resolution. The elevation distribution of nitrate well locations can be seen in Figure 12f.

Soil region The classification of soil regions was based on the Landelijk Meetnet effecten Mestbeleid (LMM) dataset, a national monitoring network established to evaluate the environmental effects of the Dutch manure policy. The LMM dataset established soil regions across the Netherlands according to factors such as dominant soil type, hydrology, and land use. The resultant soil regions can be seen in Figure 2a and nitrate distribution across soil region in Figure 2b.

Soil properties For this study, such soil-related attributes were included as bulk density (see Figure 12g), soil acidity (see Figure 12h), C:N molar ratio (see Figure 12i), calcic content (see Figure 12j), Fe-dith content (see Figure 12k), loam content (see Figure 12l), sand particle diameter (see Figure 12m), and silt content (see Figure 12n).

The properties were derived from the Soil Map of the Netherlands (Bodemkaart van Nederland) [54], which is based on the BIS-4D digital soil dataset [63], both of which are developed under the Basisregistratie Ondergrond (BRO) framework and released in 2024. These datasets integrate field sampling (systematic drillings, laboratory analyses, and expert classification) with geospatial modeling and machine learning, resulting in a 25-meter spatial resolution map. For each variable, values were assigned based on the top 0–8 cm soil layer.

Given the 16-year time frame of this study, the static map was used to represent nitrate levels throughout the entire period. However, it is sufficient because most soil properties change over multi-decadal timescales (20–50 years or more) [64]. More dynamic properties like organic matter content, pH, and C:N ratio typically vary slowly over 10–20 years under stable land use and management [65].



(a) Groundwater table (b) Precipitation (c) Temperature (d) Nitrogen deposition

(e) Population density (f) Elevation (g) Bulk density (h) Soil acidity.

(i) C:N molar ratio (j) Calcic content (k) Fe-dith content (l) Loam content in the soil

(m) Sand particle diameter (n) Silt content in the soil

Figure 12: Boxplots of spatial and environmental predictor variables used in nitrate concentration modeling.

A.6. Data preprocessing

This appendix covers the steps involved in data preparation for modeling. It covers such steps as handling missing values, feature engineering, spatial and temporal alignment, train-test split, and scaling.

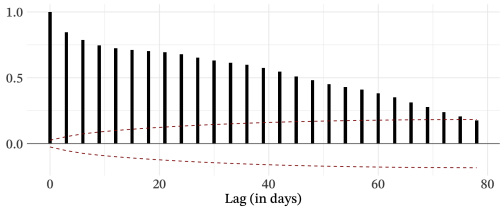
First, missing values were handled. For the target variable, all rows with missing values were removed, as they appeared to be missing at random. Nitrate levels are irregularly sampled throughout the year, making substantial gaps between measurements (usually several months). However, for the predictor variables, the issue of missing values was resolved using a combination of resampling, imputation, or removal, depending on the data source and measurement frequency. For details, refer to Appendices

A.4 and A.5.

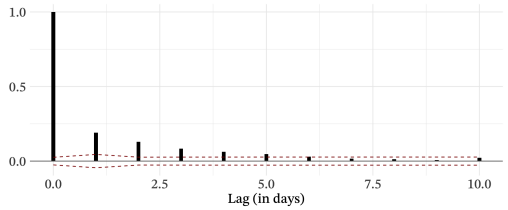
Second, feature engineering steps were performed. It has been shown that variables such as temperature, precipitation, and groundwater table have a long-term impact on the nitrate leaching process [66, 67]. Therefore, these variables were aggregated by a fixed window size, which was uniquely defined for each variable through an autocorrelation function (ACF). It is a statistical tool that measures the strength of the relationship between current values in a time series and past values at various time lags. This way, the patterns in data over time can be identified. The following equation gives the function:

$$\rho_k(x_t) = \frac{\sum_{t=1}^{N-k}(x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^{N}(x_t - \bar{x})^2}$$

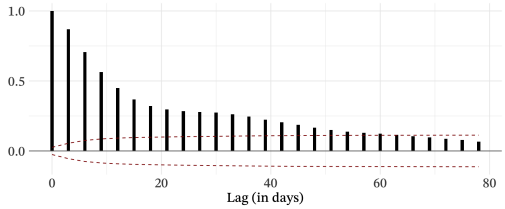
where, ρ_k is the autocorrelation at lag k, x_t is the value of the time series at time t, $\bar{x}$ is the mean of the time series, and N is the total number of observations. This results in a value between -1 and 1 that indicates the strength and direction of correlation at lag k.



(a) Autocorrelation of daily temperature. A positive correlation is observed up to 73 lags.



(b) Autocorrelation of daily precipitation. A positive correlation is observed up to 4 lags.



(c) Autocorrelation of daily groundwater. A positive correlation is observed up to 55 lags.

Figure 13: Autocorrelation plots of daily (a) temperature, (b) daily precipitation, and (c) groundwater table time series values. The dashed line represents 95% confidence bounds.

Based on ACF plots in Figure 13, the following window sizes were determined: (1) Temperature: 73 days; (2) Precipitation: 4 days; and (3) Groundwater table: 55 days.

Third, data alignment was performed both temporally and spatially. It should be noted that a spatial mismatch happened between the nitrate and groundwater table datasets, meaning that the measurement locations for nitrate and groundwater depth did not always coincide. To resolve this, each nitrate observation was paired with groundwater depth data from the nearest well within a 5 km radius. This choice was based on observed correlations between groundwater levels and distance (see Appendix A.4 Figure 11). A similar strategy was applied for weather data: temperature and precipitation. Values for these variables were taken from the closest weather station within a radius of 20 km of the nitrate monitoring site.

Fourth, while the majority of groundwater samples have low nitrate levels, there are several outliers with high concentrations. The small number of such extreme cases may limit model generalization, as it will have fewer examples to learn from when estimating these rare cases [68]. Therefore, samples with nitrate levels higher than 58 mg/L were removed from the dataset, resulting in 20 outlier data points.

Fifth, the dataset was split chronologically into training and testing sets. The training set consists of 4,439 samples spanning 13 years (2008–2020), while the testing set contains 882 samples from the last 3 years (2021–2023), corresponding to a split of approximately 83/17.

Sixth, data transformations were applied. Categorical variables were one-hot encoded to avoid ranking or ordinal assumptions between them. Continuous variables were normalized using *Min-Max*

Scaler, bringing them to a [0,1] range using the following equation:

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}},$$

where x represents an individual raw value from a given feature in the dataset. $x_{\min}$ and $x_{\max}$ denote the minimum and maximum values observed in that feature in the training set, respectively.

Once the entire dataset has undergone all the preprocessing steps, it is ready for use in model training and evaluation, which are described in the following sections.

To generate a map of estimated nitrate levels across years, a 500 x 500 m grid covering the country is created, excluding urban areas. Each grid cell was matched with input features consistent with the training data.

B. Training insights

B.1. Hyperparamter values

Hyperparameter	*Search space*	*Selected value*
Selector estimator alpha (α)	{0.1, 1.0, 10, 20, 30}	20
Max features for selector	{10, 20, 30, 40}	40
Ridge regression alpha (α)	{0.1, 1.0, 10.0, 15.0}	10.0

Table 4
Ridge Linear Regression hyperparameter search space and selected values.

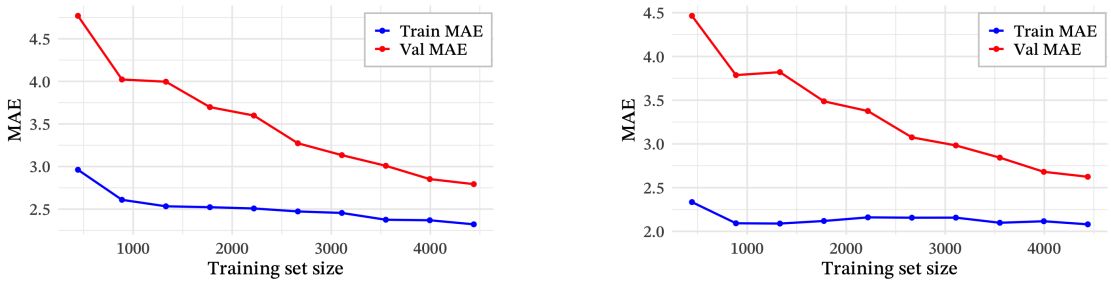
Hyperparameter	*Search space*	*Selected value*
Number of trees (*n_estimators*)	{100, 150, 200, 250}	250
Maximum depth (*max_depth*)	{None, 5, 10, 15}	None
Minimum samples split (*min_samples_split*)	{2, 4, 6, 8}	2
Minimum samples per leaf (*min_samples_leaf*)	{1, 2, 3, 4}	1
Maximum features (*max_features*)	{`sqrt`, 0.5, 1}	`sqrt`

Table 5
Random Forest hyperparameter search space and selected values

Hyperparameter	*Search space*	*Selected value*
Number of trees (*n_estimators*)	{50, 100, 150, 200, 250}	150
Maximum depth (*max_depth*)	{3, 4, 6, 10, 15, 20}	20
Learning rate	{0.01, 0.05, 0.1}	0.05
Subsample ratio	{0.4, 0.5, 0.6, 0.8}	0.8
Column sample by tree	{0.4, 0.6, 0.8}	0.8
L1 regularization (α)	{0.1, 0.5, 0.7}	0.7
L2 regularization (λ)	{2, 3, 5, 8}	5

Table 6
XGBoost hyperparameter search space and selected values.

B.2. Learning curves



(a) Random Forest learning curve

(b) XGBoost learning curve

Figure 14: Learning curves for Random Forest, XGBoost, and Ridge Linear Regression.