Contents lists available at ScienceDirect

Cleaner Water

journal homepage: www.sciencedirect.com/journal/cleaner-water

Long-term AI prediction of ammonium levels in rivers using transformer and ensemble models

Ali J. Ali [*], Ashraf A. Ahmed [*]

Department of Civil and Environmental Engineering, Brunel University London, Uxbridge UB8 3PH, UK

ARTICLE INFO

Keywords:
Transformer model
Extreme gradient boosting
Random forest regressor
River ecosystem
Toxicity
Ammonium

ABSTRACT

This study provides a cutting-edge machine learning approach to forecast ammonium (NH_4^+) levels in River Lee London. Ammonium concentrations were predicted over several time intervals using a complete dataset that includes temperature, turbidity, chlorophyll, dissolved oxygen, conductivity, and pH. Our technique captures the intricate connections between environmental conditions and ammonium concentrations using developed algorithms, including Temporal Fusion Transformer (TFT), Random Forest (RF) and Extreme Gradient Boosting (XGBoost) levels versus the important factors, considerably improving prediction accuracy. The novel aspect of this study is the utilisation of the TFT model for multi-horizon forecasting, which offers high accuracy and interpretability in hydrological predictions by combining convolutional components with an attention mechanism. The study also demonstrates the effectiveness of the TFT model in capturing short-term fluctuations while retaining accuracy over long time periods, which is a major difficulty in environmental modelling. The models used, have exceptional forecasting skills, predicting 150, 200, 365, 730, and 1095 days based on daily average and 12, 24 and 30 months based on monthly average. This dual-scale model combines flexibility and resilience, making it an effective tool for forecasting both short- and long-term environmental changes. The RF model excelled in long-term forecasts, sustaining high R-squared (R^2) (0.97) values and low root mean square error (RMSE) (0.18), and the second best one was the XGBoost with optimiser with R^2 of (0.92) and RMSE of (0.25) with forecasting 1095 days. The results also found that whilst the TFT captured the fluctuations in the short-term, it struggled with the longer-term predictions due to data granularity. The XGBoost model did remarkably well in monthly forecasts up to 12 months, maintaining low RSME. The findings also highlight the necessity of proactive water management techniques to reduce the risk of potential ecological effects, including hypoxia and oxygen depletion. The findings support resource managers in addressing prospective ammonium toxicity concerns such as oxygen depletion and ecological stress.

1. Introduction

Rivers provide several functions, such as maintaining hydrological equilibrium, facilitating floods, source of fresh water and providing food for numerous living forms. Vega et al. (1998) emphasise the need to evaluate and assess river health to maintain an ecological balance, improve water quality and fulfil domestic demand. Ammonium (NH_4^+), an essential nutrient for aquatic ecosystems, can be harmful in excessive amounts, causing eutrophication and negatively changing the quality and dynamics of the ecosystem (Huang et al., 2018). Eutrophication is the process by which water bodies become extremely nutrient-rich, resulting in excessive algae growth and oxygen depletion, is a severe hazard to aquatic life (Akinnawo, 2023). Ammonium contamination comes from various sources, including diffuse agricultural runoff and point sources like urban and industrial water discharge (Maranon et al., 2006; Krapac et al., 2002). Therefore, targeted strategies and focused methods for reducing ammonium concentrations in aquatic environments are critical as their conversion under certain conditions poses an indirect yet potential risk to human health (Britto and Kronzucker, 2002; Lin et al., 2017). These types of health conditions vary from liver diseases, kidney diseases, immune system problems and encephalopathy (Kanjilal et al., 2024) Although ammonium (NH_4^+) is less toxic than Ammonia (NH_3) (Sawyer, 2008), but still it plays a crucial role in the nitrogen cycle and the overall aquatic ecosystem health (Bhatnagar and Sillanpää, 2011). The equilibrium between ammonium (NH_4^+) and the unionised ammonia (NH_3) is determined by water pH and temperature,

[*] Corresponding authors.
E-mail addresses: ali.ali@brunel.ac.uk (A.J. Ali), ashraf.ahmed@brunel.ac.uk (A.A. Ahmed).

https://doi.org/10.1016/j.clwat.2024.100051
Received 12 August 2024; Received in revised form 5 October 2024; Accepted 19 October 2024
Available online 24 October 2024
2950-2632/© 2024 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

which can vary greatly in natural water bodies (Nollet and De-Gelder, 2000). As the pH in water increases and the surrounding conditions become more alkaline, ammonium converts to ammonia, which is especially harmful to aquatic life. This accelerated transition takes place at higher temperatures, making warm, alkaline waters more sensitive to ammonia toxicity. Although ammonium dynamics in wastewater (Icke et al., 2020) and groundwater (Perović et al., 2021) have been the subject of several research, a limited focus has been provided to accurately predict ammonium levels in river systems, especially over long horizon where varying environmental conditions plays critical conditions. These extended horizons often introduce complexity that conventional models frequently fail to account for. The complex non-linear interactions and temporal dependencies in river ecosystems are overlooked by statistical models and simple machine learning approaches, leading to predictions that become unreliable as environmental circumstances change over time.

Machine learning approaches have been used in several research to simulate the ammonium levels in river ecosystems. For example, Khullar and Singh (2021) offered a thorough analysis of machine learning models for predicting parameters like dissolved oxygen (DO), chemical oxygen demand, and biochemical oxygen demand, such as artificial neural network (ANN) adaptive neuro-fuzzy inference systems (ANFIS), and support vector machines (SVM). Even while these models showed a great deal of promise in capturing the intricate non-linear correlations between the input variables, they frequently struggle to handle temporal dependencies and provide forecasts with many horizons. However, most of these models are designed mainly for predictions over a single time horizon, and they are not robust enough to be employed in long-term forecasting scenarios where environmental factors are changing over time. They also frequently need a lot of data for training and are sensitive to the quality of the data, which presents a problem in real-world applications where the data is frequently irregular or sparse.

1.1. *Amounts of ammonium in water*

According to the Environment Agency (2014), in unpolluted waterways, ammonium concentrations usually fall between 0.2 and 1 mg/l, with a focus on the nitrogen (N) component of the compound. However, levels in treated sewage effluent can vary between 10 and 20 mg/l as N. It is noteworthy that unionised ammonia concentrations as low as 0.025 mg/l may harm fish (Parvathy et al., 2023), demonstrating the narrow margin between acceptable and hazardous levels. An ammonium concentration of 2.5 mg/l as NH_4^+ indicates potentially dangerous circumstances for aquatic life, emphasising the significance of constant monitoring and control of water quality (Chapman, 1996). whilst the dangers of ammonium pollution are generally acknowledged, current research has mostly focused on groundwater and wastewater sources of contamination, with rivers receiving far less attention. Notably, while studies by Liang et al. (2022), Zhang et al. (2020) and Ayejoto et al. (2022) have investigated ammonium dynamics in groundwater and wastewater, respectively, and others, such as Covatti and Grischek (2021), have touched on river systems, the specific focus on river water, particularly in the context of the United Kingdom, remains unexplored. This gap persists despite periodic examinations of riverbank implications on pollution levels (Groeschke et al., 2017; De Vet et al., 2010).

1.2. *Ammonium and aquatic life*

Understanding and predicting NH_4^+ concentrations in river ecosystems is crucial. This is because it not only involves managing nutrients, but also involves anticipating conditions that can lead to the formation of hazardous ammonia. High NH_3 levels can harm fish and invertebrates by disrupting their respiratory and reproductive systems. This can result in respiratory problems, toxin build-up, and even death (Azrour et al., 2022). NH_4^+ can also combine with other molecules in water to form nitrites and nitrates, which can be harmful to human and animal health (Mejía and Barrios, 2023).

1.3. *Traditional techniques for measuring* $\boldsymbol{NH_4^+}$

Conventional techniques like the Nessler method, evaporation determination method, indicator method, and fluorescence method for ammonia nitrogen detection suffer from complex procedures, low sensitivity, and limited accuracy (Wang et al., 2023). Indophenol blue colourimetry is one of the most famous methods for measuring ammonium concentrations. However, this method has considerable limitations due to its complexity, toxicity and time-consuming (Ma et al., 2018; Holmes et al., 1999). Additionally, real-time monitoring is a viable avenue for forecasting concentrations, but its high cost, as well as the limitations of standard approaches to capture nonlinear and nonstationary water quality data, highlight the need for a more advanced methodology (Li et al., 2022). The limitations of existing monitoring approaches, which frequently fall short of sensitivity, accuracy, and practicality for real-time analysis, highlight the need for a paradigm change.

Existing models, including linear regression and basic decision tree methods, often lack the ability to capture the complex, non-linear correlations and long-term dependencies that are present in environmental data (Maganathan et al., 2020). Due to these constraints, multi-horizon forecasts, which are essential for efficient water management and policy formulation, perform poorly

1.4. *Introduction to machine learning*

The combination of these limitations highlights the urgent need for a revolutionary solution. Machine learning (ML) emerges as a cutting-edge alternative that has the potential to revolutionise this measurement with remarkable efficiency and precision. Looking further into ML, some models stand out for their novel methods of predictive analysis in environmental research. In this study, we employed models like Temporal Fusion Transformer (TFT), XGBoost and Random Forest Regressor, which are known for productive accuracy and adaptability to various data types

TFT utilises a novel architecture combining convolutional components with an attention mechanism for multi-horizon forecasting, as Lim et al. (2021) outlined. It provides a unique mix of flexibility and interpretability, making it especially useful for hydrological prediction, as noted by Fayer et al. (2023), Ahmed et al. (2024), and evidenced by Ali et al. (2024). In the context of ammonium concentration in rivers, TFT excels in forecasting by effortlessly integrating numerous data sources and handling complexity with the requirement of hyperparameter modifications (Lim et al., 2021). This capability is significant as it allows the model to learn and adjust to changes in ammonium levels over time. The versatility of this makes it perfect for studying riverine ecosystems, where comprehending the intricate variability and interactions of multiple variables, particularly those affecting ammonium levels, is crucial.

Following TFT's unique approach, Extreme Gradient Boosting (XGBoost) stands out for its remarkable efficiency and performance in classification and regression problems. XGBoost, developed by Chen and Guestrin (2016), uses a clever method to optimise both speed and processing resources while maintaining model performance. XGBoost's resistance to overfitting, achieved mostly through built-in regularisation algorithms, makes it especially useful in environmental data analysis, where avoiding overly complicated models is critical (Wang and Ni, 2019). Furthermore, XGBoost model feature significance evaluation identifies the most important predictors of ammonium levels, allowing researchers to pinpoint crucial environmental elements impacting water quality. This feature not only improves the model's interpretability, but it also informs future data gathering and policy-making efforts. A comprehensive review of the different ensemble methods models for hydrological, including river water quality predictions, is provided by Zounemat-Kermani et al. (2021)

In addition to the complex techniques of TFT and XGBoost, the Random Forest Regressor appears as a critical component in the hydrology field (Tyralis et al., 2019). Random Forest (RF) was developed by Breiman (2001), and was considered one of the most successful machine learning algorithms in water science and hydrological applications (Tyralis et al., 2019). It has been applied in various scientific areas, including agriculture (Liakos et al., 2018), land cover classification (Gislason et al., 2006) and biological studies (Goldstein et al., 2011). RF, which is based on the ensemble learning paradigm, uses several decision trees to create a more generalised model, considerably lowering the risk of overfitting (Breiman, 2001). Its strength is the ensemble method, aggregating predictions from several trees to improve forecast accuracy and stability. RF's capacity to generate estimates of feature value is consistent with our goal of discovering significant predictors of ammonium content in rivers.

1.5. Model performance metrics and validation techniques

Regarding model evaluations, it is critical to address the methods to measure their performance in time-series forecasting. Rolling window analysis has historically been employed for such evaluations, giving information on a model's stability across time (Kombo et al., 2020; Hussein et al., 2020). This approach computes parameter estimates throughout a fixed-size window of the sample, providing a measure of parameter constancy, which is critical in dynamic scenarios. However, using this approach in hydrological forecasting, particularly to estimate ammonium levels in rivers, necessitates subtle customisation. Riverine ecosystems, unlike groundwater systems, have significant temporal and geographical variability as a result of their direct interaction with both land and atmospheric systems (Dingman, 2015). This interaction provides a degree of complexity and unpredictability that challenges the assumption of constant parameters, a topic frequently discussed in the context of financial time series (Zivot et al., 2003). To improve the robustness of our evaluation, we also used the holdout approach (Roelofs et al., 2019; Cerqueira et al., 2020). This strategy supports rolling window analysis by reserving a portion of the dataset for final assessment, guaranteeing that the model's performance is evaluated against previously unknown data. Such an approach is required for a thorough knowledge of the model's forecasting powers and flexibility to the inherent diversity seen in river ecosystems.

1.6. Research gaps and novelty

This research marks a breakthrough in environmental science by combining hydrological knowledge with cutting-edge machine learning tools to accurately anticipate ammonium (NH_4^+) concentrations in river ecosystem. Key innovations in this study includes:

- This analysis goes beyond traditional methods by incorporating a dynamics assessment of NH_4^+ levels against environmental factors such as temperature, pH, turbidity, chlorophyll, dissolved oxygen, and conductivity. This helps predict conditions that may lead to ammonia toxicity in river ecosystems.
- The new application of advanced machine learning models, such as the Temporal Fusion Transformer (TFT), a strong tool for dealing with time-series data, in conjunction with well-established approaches such as Extreme Gradient Boosting (XGBoost) and Random Forest Regressor (RF). This combination takes advantage of the benefits of both cutting-edge and classic algorithms to improve prediction performance.
- The models established in this work can anticipate ammonium levels over lengthy periods of time utilising daily and monthly data. They are highly accurate, forecasting up to 30 months ahead with monthly averages and up to 3 years with daily data. This dual-scale method is flexible and resilient in anticipating both short- and long-term environmental changes, making it an effective tool for environmental management.
- The study fills the gap between theoretical research and actual application by providing a scalable and adaptable paradigm for river systems beyond the United Kingdom. This versatility helps to promote global water sustainability by offering a useful resource for regulating water quality in a variety of environmental scenarios.

This study offers a significant advancement in predicting ammonium levels in river ecosystem and by incorporating advanced machine learning models such as TFT, XGBoost and Random Forest. The study's dual-scale forecasting capacity enhances its applicability, offering a valuable tool for water resource managers to reduce ecological hazards related to ammonium toity.

2. Materials and methods

This section outlines our approach for estimating ammonium levels in a UK river. It uses a combination of deep learning and decision tree models, including TFT, XGBoost, its enhanced variant with random search optimisation, and Random Forest Regressor. We describe our data collecting and preparation methods, discuss the mechanics of each ML model, and explain the assessment criteria.

2.1. Data collection and pre-processing

We compiled large historical data from River Lee located in the heart of central London, UK's capital city. The data source is from the Environment Agency's Hydrology Data Explorer, which provides extensive spatial and temporal hydrological data across the United Kingdom. The area of interest was a stretch of the River Lee in East London, England, which is about 7 kilometres from the River Thames in East London. Data on important water quality parameters, including ammonium levels, temperature, turbidity, chlorophyll, dissolved oxygen, conductivity, and pH, were supplied by a number of monitoring sites throughout the country. However, the selection of the monitoring station was based on the availability of these parameters, as not every station had full data on all features, and some were newly recorded.

The data selected were from March 2016 to January 2024. These data were used to predict ammonium NH_4^+ levels in river ecosystems by refining data pre-processing procedures to better reflect the complexities of hydrological and nitrogen cycles (Yang et al., 2021). Although the model's dependence on historical data gives it a strong basis, changes in the environment, such as shifting climatic patterns, urbanisation, or policy changes, could have an impact on the model's accuracy in the future. These changes may affect the relationship between environmental factors and ammonium levels, which could impact forecast accuracy. Regular retraining with current data could help alleviate issue and keep the model relevant. Moreover, adding scenario-based techniques and outside data sources, such climate projections, might enhance the model's ability to include future uncertainty. (Newhart et al., 2020)

Initially, the data was collected at hourly intervals, which resulted in some temporal coverage irregularities. As a result, the normalisation approach was applied to a uniform daily time series (Kang and Tian, 2018), with the mean value of available data calculated for each day. We efficiently minimised the effect of missing records by calculating the mean values for each day, ensuring that our datasets mirrored the daily resolutions, which provided a better assessment of hydrological trends and nutrient cycling patterns within the aquatic system. This method of using daily averages is especially relevant to our knowledge of the nitrogen cycle. This step is critical for models like TFT, XGBoost, and Random Forest, which are sensitive to the scale of input features (Vafaei et al., 2018; Murray et al., 2010). Our aim is to improve model convergence, stabilise the learning process and stop any single feature from dominating due to scale disparities by normalising the data. Normalisation can accelerate the convergence of deep learning models such as TFT by balancing the gradients and preventing problems with disappearing or bursting gradients (Wu et al., 2020b). Additionally, normalisation contributes to more regular and understandable splits, which enhances the accuracy and generalisability of tree-based models like Random Forest and XGBoost. The monthly analysis followed the same procedure. We computed the mean values of the daily data for each month to produce a consistent monthly time series. This method guaranteed that the datasets ha a consistent temporal resolution, allowing for a full analysis of long-term hydrological and nutrient cycle patterns. By using standard approaches on both daily and monthly datasets, we guaranteed that our prediction models were reliable and comparable across multiple time periods. Since changes in ammonium levels are driven by both biotic and abiotic processes, that was explained by Yang et al. (2021) on a scale that hourly data cannot effectively reflect. Furthermore, a linear interpolation (Huang, 2021) approach was implemented to retain the temporal integrity for analysing quality trends. This technique preserved the dataset's integrity and enabled continual examination of environmental factors influencing NH_4^+ concentrations. Although this technique offers a straightforward and efficient way to generate continuous time series data, it assumes that values will transition smoothly, which may not capture the sharp fluctuation or abrupt swings that might happen in dynamic riverine environments (Liu et al., 2011). These biases can affect machine learning like TFT and XGBoost, which depend on precise input data to identify patterns and make accurate forecasts more sensitive. In order to reduce this risk, the influence of interpolation on model accuracy was minimised by utilising a holdout dataset with low interpolated values to test the model's performance (Cerqueira et al., 2020).

Daily and monthly durations were chosen to properly capture ammonium dynamics since they reflect the impact of both short-term and long-term processes on ammonium concentrations. Daily timescales enable the models to detect quick variations in ammonium concentrations induced by diurnal cycles, wastewater discharge, or rainfall events, which is crucial for real-time management. Monthly timeframes, however, aid in detecting seasonal tendencies. This approach is consistent with Watson et al. (2017) findings, who demonstrated that short-term scales such as daily to weekly intervals are critical for understanding variability in environmental factors such as rainfall, which has a direct impact on nutrient concentrations in water bodies (Watson et al., 2017).

The pre-processing steps are to maintain the dataset's homogeneity and correspond with our goal of understanding environmental influences on river ecosystems (Wang et al., 2023). By fine-tuning our dataset to represent the daily cycles of hydrological and nitrogen processes, we enabled a more thorough and contextually appropriate study with advanced machine learning algorithms. In addition, it provides us a vital insight into the temporal dynamics of river ecosystems and a better understanding of water quality management.

We combined a wide range of environmental parameters, such as pH, temperature, dissolved oxygen, turbidity, chlorophyll and conductivity, with ammonium levels to create a multidimensional dataset targeted at forecasting future ammonium concentration across a variety of time periods. This approach ensures that our predictive models capture the complex processes that influence ammonium levels in the river environment, as was done by Li and Li (2023) to predict ammonia nitrogen.

Furthermore, we methodically proceeded to split the dataset into training, validation and holdout purposes. This data splitting technique, designed to thoroughly analyse the model's performance (Ransom et al., 2017), supports our aim of predicting accurate and actionable forecasts for successful water quality management. Allocating 70 % of the data to the training set enables the models to get a thorough grasp of the interactions between numerous environmental conditions and ammonium levels, which is essential given the complexities of these natural processes. This extensive training volume allows the models to recreate the subtle patterns found in river ecosystems properly. The remaining data was distributed evenly between validation and holdout sets, with each getting 15 %. The validation set is critical throughout the model tuning phase, allowing adjustments to improve forecast accuracy whilst avoiding overfitting. Conversely, the 15 % holdout set, designated for the final evaluation, serves as a rigorous test of the model's capacity to generalise to new, unseen data, reflecting standard ML practices for ensuring the final assessment is both strict and reliable (Dwork et al., 2015).

It is important to note that despite these quality assurance measures, limitations in the data sets should be acknowledged. For instance, because of the monitoring stations' poor spatial resolution, certain areas of the river were not as well-represented as others. Furthermore, even though the dataset had sufficient temporal coverage, unforeseen weather occurrences or abrupt changes in land use during the research period might have introduced unpredictability that the data did not completely capture.

2.2. *Machine learning models*

2.2.1. *Temporal fusion Transformer (TFT)*

Our methodology for predicting ammonium (NH_4^+) levels in river ecosystems incorporates advanced ML techniques and uses the TFT, which has an unsurpassed capacity to handle complicated, multi-temporal data (Marcellino et al., 2006; Li et al., 2019). To enrich the input data with temporal dynamics indicative of hydrological cycles and nitrogen transformations, pre-processing steps included building a dataset with lagged and rolling features for numerous environmental parameters, with exception of target variable (NH_4^+) (González-Enrique et al., 2022). We specifically structured the dataset to contain previous observations (lagged features) and smoothed trends (rolling averages) across defined time periods, corresponding to natural changes in river water quality and the interaction of causes controlling ammonium levels.

The TFT model, designed for the sensitive problem of ammonium prediction, has a sophisticated attention-based architecture that can deconstruct our dataset's deep temporal correlations (Vaswani et al., 2017). This design enables a thorough knowledge of how temperature, pH dissolved oxygen and other factors interact over time to alter concentrations. The architecture of the model includes components such as a gating mechanism and variable selection networks, which are capable of filtering through noise to focus on the most important predictors at any given time (Clevert et al., 2015; Appels et al., 2015; Dauphin et al., 2017). Such accuracy is essential in hydrology, where many causes might affect water quality differently (Kushwaha et al., 2024).

The Gated Residual Network (GRN) and Variable Selection Networks are critical components of the TFT model's operation, allowing for selective processing and prioritisation of input data (Lim et al., 2021):

- The GRN uses Layer Normalisation and Gated Linear Units to adapt dynamically to changing environmental factors, improving model accuracy and efficiency. The formula for the GRN operation is:

$$GRN_{\omega}(a, \quad c) = LayerNorm(a + GLU_{\omega}(\eta_1))$$

$$\eta_1 = W_1\omega\eta_2 + b_1\omega$$

$$\eta_2 = ELU(W_2\omega a + W_3\omega c + b_2\omega)$$

- Variable Selection Networks focus the model's attention on the most relevant variables at each timestep, which is crucial for regulating the many impacts on NH_4^+levels. The mechanism works as follows:

$$\upsilon_{Xt} = Softmax(GRN_{\upsilon X}(\Xi t, cs))$$

The implementation begins with altering the scaled features using lagging and rolling, followed by dividing the dataset into training, validation, and holdout sets to ensure a thorough assessment framework. The TFT model was built using embedding layers for input processing, multi-head self-attention layers for capturing long-term dependencies, and LSTM layers for temporal pattern recognition. It ended with a dense output layer for exact amplitude level forecasting. The training was optimised with an early-stopping call-back to prevent overfitting and keep the model generalisable to new data. TFT was selected for its capacity to capture long-term temporal dependencies and manage different time horizons, which makes it an ideal model for non-linear and time-varying interactions in the River Lee dataset.

2.2.2. *Extreme gradient boosting (XGBoost)*

Building on the approach used for the TFT, we expand our prediction framework by incorporating XGBoost, a scalable ML system known for its success in tree boosting (Chen and Guestrin, 2016). This technique stands out for its scalability and has been widely used by data scientists to generate cutting-edge outcomes across various ML. XGBoost presents a unique sparsity aware algorithm, which considers data independently for approximation tree learning, making it ideal for our study's focus on forecasting ammonium levels in river ecosystems (Nalluri et al., 2020).

XGBoost's capacity to handle sparse data is critical for environmental datasets, which frequently contain missing values or zero entries as a result of data collection and feature engineering techniques such as one-hot encoding (Abou Omar, 2018). By adding a default direction for tree nodes in the case of missing data and using a sparsity-aware split finding technique, XGBoost ensures that all accessible data contributes to the model's learning process, no matter how partial. Given the unpredictability and occasional gaps in environmental monitoring data, this is a very important consideration for our research. XGBoost optimises the following objective function, which is critical for understanding how effective it is in handling complex environmental datasets (Chen and Guestrain, 2016):

$$Obv(\theta) = \sum\nolimits_{i=1}^{n} l\left(\mathscr{Y}_i, \quad \widehat{\mathscr{Y}}_i^{(t)}\right) + \sum\nolimits_{k=1}^{t} \Omega(f_k)$$

Where, l is the differentiable convex loss function that computes the difference between the expected $\widehat{\mathscr{Y}}_i^{(t)}$ and the actual $\mathscr{Y}_i$ values. Furthermore, the regularisation term is denoted by Ω with $f_k$representing the k-th tree. The regularisation term smooths the final learning weight to prevent overfitting. This is especially important since some of the variables with the ammonium do not have a direct impact on ammonium levels, such as chlorophyll and conductivity. Additionally, XGBoost's distinct contribution to gradient boosting is the efficient implementation of tree learning methods and the regularisation term, which is defined as (Chen and Guestrain, 2016):

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum\nolimits_{j=1}^{T} w_j^2$$

Where, T is the number of leaves in the tree, $\gamma and \lambda$ are the parameters that control the complexity of the model and w_j is the score on the j-th leaf. XGBoost was selected because of its resilience and capacity to avoid overfitting, particularly in big and complex datasets. It is well-known for managing non-linear connection and missing data.

2.2.3. *XGBoost with random search optimization*

We further enhanced our model's efficacy by using Random search optimisation for hyperparameter tuning that was developed by Bergstra and Bengio (2012). This optimisation approach is critical for finetuning XGBoost's complex parameters, ensuing that the model operates not only accurately but also efficiently across a wide range of environmental conditions. Random search (RS) is used to systematically explore the model's enormous hyperparameter space, which includes a wide range of parameters (Haidar et al., 2019). In this study, we used learning rate, number of trees, tree depth, regularisation, sum of instance weight, and subsample ratio. Random search, unlike grid search, which thoroughly evaluates all potential hyperparameter combinations, randomly selects a selection of parameter combinations. This technique is both time-efficient and frequently produces comparable high performance, making it ideal for high-dimensional hyperparameter spaces (Bergstra and Bengio, 2012; Putatunda and Rama, 2019). The process involves specifying a distribution for each hyperparameter rather than a discrete set of values, resulting in a greater search range and the ability to reveal the optimum settings that would otherwise be ignored by a more rigid, grid-like search algorithm. It is worth mentioning that the hyperparameter approach did not include cross-validation because the main focus was on employing a holdout set for final validation. Although overfitting is often prevented via cross-validation, we employed a separate validation set to track the model's performance and avoid overfitting during tuning, and a holdout test set for the final model assessment. This approach avoided the additional computational burden of cross-validation and guaranteed a trustworthy evaluation of model generalisability (Cerqueira et al., 2020).

Each iteration of the random search selects a set of parameters, which are then assessed by training the XGBoost model and testing its performance on a validation set (Bergstra and Bengio, 2012). In our study, RMSE (Root Mean Squared Error) is the performance parameter used to steer optimisation, providing a clear standard for model accuracy in forecasting ammonium levels. This hyperparameter was for improving performance through the use of random search optimisation, which decreased overfitting and increased accuracy over a range of time horizons.

2.2.4. *Random forest regressor*

We also incorporate the Random Forest Regressor (RF) to make use of its powerful ensemble learning capabilities. Random Forests, developed by Breiman (2001), uses numerous decision trees to improve forecast accuracy and reduce model overfitting, making them especially useful for complicated environmental datasets. Each tree in the RF is constructed using a bootstrap sample, which is a randomly selected subset of the retraining data (Biau and Scornet, 2016). This process is called bagging or bootstrap aggregating, which is a procedure that helps to reduce variation. Furthermore, at each node of the tree, a subset of characteristics is randomly chosen to decide the split, which adds randomisation to the trees and makes the model more robust to noise. This sampling method can be expressed as (Biau and Scornet, 2016):

$$D_i^* \sim Uniform(D), fori = 1, \ldots, n$$

Where D is the dataset and n is the number of samples. The decision tree construction can be expressed as follows:

$$\Delta Var(t) = Var(t) - \left(Var\left(t_{left}\right) + Var\left(t_{lright}\right)\right)$$

Where, the variance of the target variable at nude t is denoted by $Var(t)$ and the nodes resulting from the split are denoted by t_{left} and t_{lright}, respectively.

Our implementation defines the RF model's hyperparameters, such as a number of trees, the depth of each tree and a minimum number of samples required. These parameters are modified using RandomizedSearchCV, which optimises for the optimal combination based on cross-validation findings and is consistent with the theoretical framework mentioned above for handling the bias-variance trade-off required in environmental modelling (Shaikh-Mohammad and Siddiqui, 2021). The selection of RF was based on its robust ensemble learning capabilities, which minimise overfitting and offer insights into feature significance. This makes RF an excellent choice for determining the primary drivers of ammonium levels in dynamic river systems.

2.3. Evaluation methods

2.3.1. Rolling window analysis

Each model in our study was assessed using a consistent assessment process geared to our goal of forecasting ammonium levels in river ecosystems. We employed a rolling window for post-data pre-processing; it generates lagged and rolling window features that help us understand how previous environmental conditions influence the current ammonium levels. The rolling window technique evaluates a model's performance at a single time instance i utilising data from $i-1$ to $i-N$ prior observations, providing h-step forward predictions (Amor et al., 2016). This strategy is consistent with the dynamic character of river habitats, in which previous events and circumstances change water quality measures like ammonium levels. As the window moves forward one period at a time, the model updates continually, adding fresh data to improve future forecasts. This continuing modification helps to reflect the constant changes observed in natural water systems, ensuring that our projections are relevant and accurate throughout time.

2.3.2. Holdout set method

The holdout method is an important methodology for measuring the performance of time series forecasting models, particularly for non-stationary time series. This strategy separates the data into two sets: training and testing. Models are trained on the first segment before being tested on the second, allowing for the evaluation of previously unknown data. This methodology is especially useful for non-stationary series since it provides more robust validation than approaches like cross-validation. The holdout approach is important because it tests the model's ability to anticipate fresh, previously unknown data, which is critical for assuring prediction dependability in real-world settings (Cerqueira et al., 2020).

2.3.3. Performance metrics (RMSE, R2, MAE and NSE)

We evaluated our models' performances using three metrics: Root Mean Square Error (RMSE), R-squared (R^2), Mean Absolute Error (MAE), and the Nash-Sutcliffe Efficiency coefficient (NSE). These measures were chosen to give a thorough assessment of the model's prediction accuracy and dependability (Chicco et al., 2021; Duc and Sawada, 2023).

- The RMSE evaluates the size of predicted mistakes, providing information about the average magnitude of the error. It takes the square root of the average squared discrepancies between predicted and actual values, which provides a clear measure of the model's accuracy. The formula for the RMSE is:

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(X_i - Y_i)^2}; (best\ value = 0 \quad , worst\ value = +\infty)$$

- MAE calculates the average magnitude of mistakes in a series of predictions without considering their direction. It's the average over the test sample of the absolute differences between prediction and actual observation, with all individual deviations having equal weight.

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|X_i - Y_i|; (best\ value = 0 \quad , worst\ value = +\infty)$$

- R-squared (R^2) measures how much of the dependent variable's variation is predicted from the independent variables. Based on the fraction of total variance explained by the model, it assesses how effectively observed outcomes are replicated by the model.

$$R^2 = 1 - \frac{\sum_{i=1}^{m}(X_i - Y_i)^2}{\sum_{i=1}^{m}(\overline{Y} - Y_i)^2}; (best\ value = +1 \quad , worst\ value = -\infty)$$

- The Nash-Sutcliffe Efficiency (NSE) is particularly important in hydrological modelling to assess model prediction ability. An efficiency of 1 (NSE = 1) indicates a perfect match between the modelled discharge and the observed data. An efficiency of 0 (NSE = 0) implies that the model predictions are as accurate as the observed data's mean, whereas an efficiency less than zero shows that the observed mean outperforms the model. It's computed as (Duc and Sawada, 2023):

$$NSE = 1 - \frac{\sum_{i=1}^{n}(\mathscr{Y}_i - \widehat{\mathscr{Y}}_i)^2}{\sum_{i=1}^{n}(\mathscr{Y}_i - \overline{\mathscr{Y}})^2}; (best\ value = +1 \quad , worst\ value = -\infty)$$

3. Results and discussion

This section is divided as follows: first, we presented the correlation heat map to present the relationships between each variable and ammonium. This was followed by the results in the training phase, noting that the value of NSE results closely matched the values of R^2 results in the holdout sets and that agrees with Duc and Sawada (2023). The daily predictions are then shown in the scatter plot, with an emphasis on the RMSE value for the holdout sets, to observe the robustness of the models on new unseen data. To give a thorough understanding of the model's performance, we also present the plots of actual against predicted data and the error distribution data. Lastly, we show the outcomes of the monthly data forecasts, pointing out that the models had a difficult time efficiently learning the patterns because the monthly averages had shrunk the data size. This section is structured to give a thorough understanding of each model's prediction capabilities by highlighting its strengths and weaknesses in relation to various time scales and data combinations.

We use powerful machine learning approaches to analyse the performance of our models over multiple time scales and configurations, improving our capacity to forecast key water quality metrics that are important for environmental management and policy-making. The data used in this study included all of the variables, which are temperature, turbidity, chlorophyll, dissolved oxygen, conductivity, and pH, with ammonium serving as the target variables. This comprehensive method guarantees that the interaction between hydrological parameters and ammonium concentrations is effectively documented, resulting in a strong foundation for forecasting situations that may lead to ammonia toxicity in river ecosystems (Li and Li, 2023).

3.1. Correlation analysis

The correlation heatmap (Fig. 1) helps understand the relationships between the variables (Ji et al., 2019), specifically between ammonium levels and the other variables. Notably, there is a moderate negative correlation (-0.16) between NH_4^+ and dissolved oxygen (O_2), indicating that greater O_2 levels often correlate to lower ammonium concentration. This also aligns with the findings of Ortiz-Santaliestra and Marco (2015). They also found this link is crucial because excessive ammonium levels can deplete oxygen, resulting in hypoxic conditions that harm aquatic life. Furthermore, a modest positive correlation of (0.17) with temperature indicates that rising water temperatures may encourage higher NH_4^+ concentration, which is a concern in the context of climate change (Jones and Hood, 1980). The weak correlation between chlorophyll, pH and turbidity indicates that these variables may impact ammonium levels indirectly or through complicated interactions. During the early stages of water bloom, increased algal density leads to increased

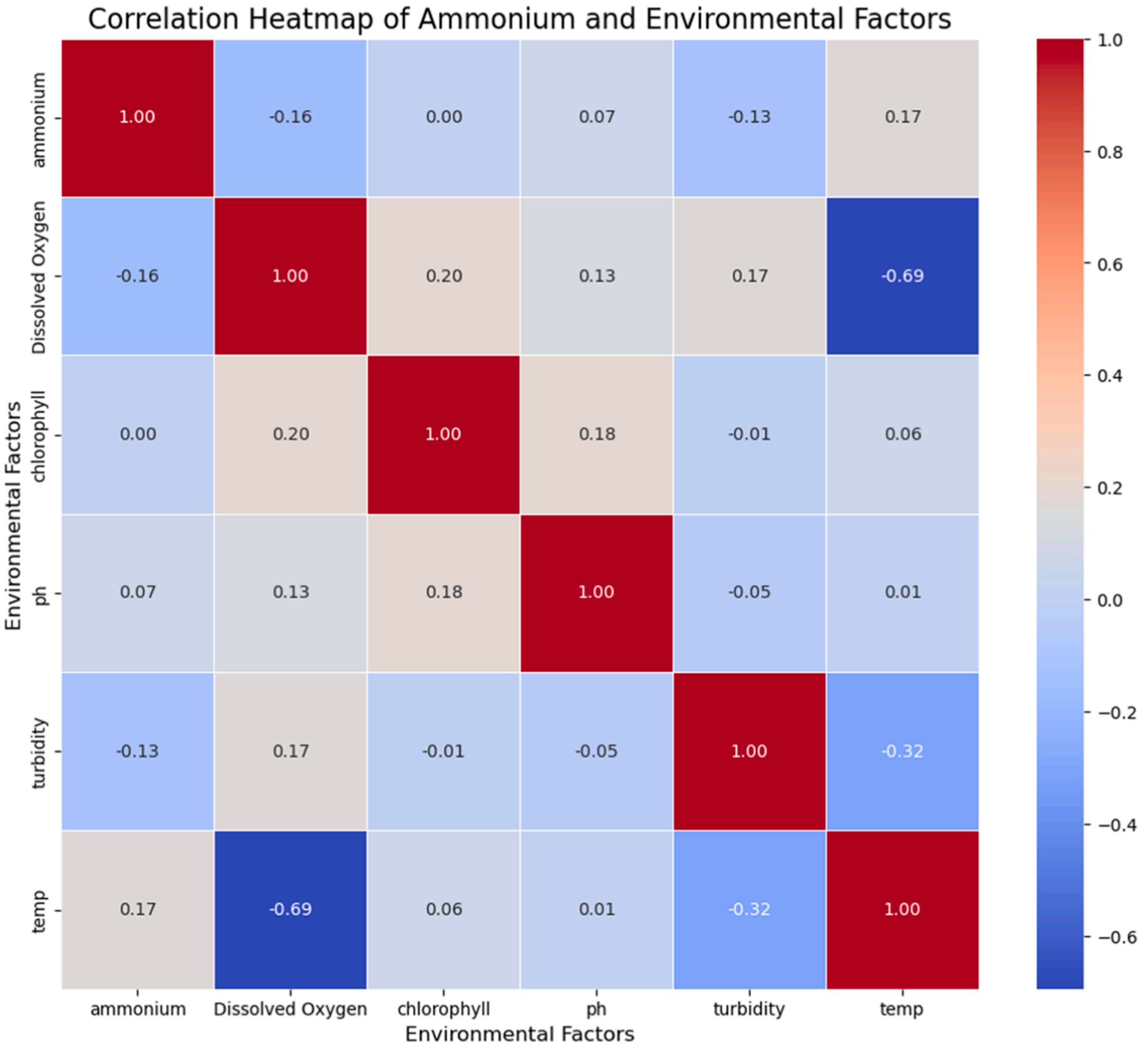

Fig. 1. Correlation of the 7 parameters.

chlorophyll levels, which improve photosynthesis and O2 levels (Kube et al. 2018). However, algae growth on the water's surface prevents sunlight from reaching organisms, resulting in aerobic respiration and fast oxygen consumption (Huang and Zhang, 2024).

From a hydrological perspective, these conclusions emphasise the interdependence of physical, chemical and biological processes in river systems. Fundamental hydrological processes include dissolved oxygen levels, which are regulated by water movement, temperature and biological activity. The substantial negative association (-0.69) between O_2 and temperature highlights the need to maintain normal flow regimes and thermal conditions to support oxygenation and manage ammonium levels. Understanding these relationships is critical for developing effective water management strategies, such as increasing riparian vegetation to regulate water temperature and improve dissolved oxygen levels (Dugdale et al., 2024) and managing nutrient inputs from agricultural and urban runoff to avoid excessive ammonium accumulation. This integrated approach is critical for forecasting and mitigating the effect of environmental changes on river ecosystem, therefore preserving their health sustainability.

3.2. Model performance

The performance of the final model was assessed on the holdout datasets to determine its generalisability. The holdout set, containing data that is not used during model training or validation, but used to test important metrics such as RMSE, MAE and R2, showing the models' prediction accuracy and generalisability.

The results in Table 1 and Table 2 present that the prediction accuracy varies dramatically between models and prediction horizons. XGBoost model and its variant with random search optimisation outperformed other models in daily data predictions, consistently reaching near-perfect R^2 values, which agree with Chicco et al. (2021) and high holdout NSE values. The TFT likewise performed well, although with somewhat higher RMSE and lower NSE values than XGBoost with random search. In the monthly data prediction, XGBoost consistently outperformed the other models, with excellent R^2 values and high holdout NSE values. However, it is clear that the models' performance degrades across longer prediction horizons, especially in the TFT model, which showed increasing RMSE and falling NSE values as the prediction horizon increased. The consistent model performance across short time periods could have been facilitated by the use of linear interpolation (Huang, 2021) for missing data. However, it could also explain some of the limitations in capturing long-term trends or sudden fluctuations, as observed by the TFT model's increasing RMSE over longer horizons.

Despite its efficacy, the TFT model may suffer to anticipate in the long run due to the data granularity (Cirillo et al., 2021). This can cause overfitting on shorter scales and underperformance on longer ones as the model's assumptions lose validity over time (Popovic et al., 2015). When the model is rich with information, which in this case the daily data, it could efficiently capture short-term fluctuations and seasonal

Table 1
Daily data.

Prediction Horizon	Model Type	Train			Holdout NSE
		RMSE	MAE	R^2	
150 Days	TFT	**0.16**	**0.10**	**0.97**	**0.900**
	XGBoost	**0.25**	**0.20**	**0.94**	**0.875**
	XGBoost (Random Search)	**0.05**	**0.04**	**1.00**	**0.924**
	Random Forest	**0.05**	**0.02**	**0.99**	**0.832**
200 Days	TFT	**0.17**	**0.11**	**0.97**	**0.877**
	XGBoost	**0.24**	**0.18**	**0.94**	**0.874**
	XGBoost (Random Search)	**0.05**	**0.03**	**1.00**	**0.905**
	Random Forest	**0.05**	**0.02**	**1.00**	**0.849**
365 Days	TFT	**0.16**	**0.10**	**0.97**	**0.894**
	XGBoost	**0.21**	**0.17**	**0.96**	**0.880**
	XGBoost (Random Search)	**0.08**	**0.06**	**0.99**	**0.914**
	Random Forest	**0.05**	**0.02**	**1.00**	**0.828**
730 Days	TFT	**0.14**	**0.09**	**0.99**	**0.913**
	XGBoost	**0.13**	**0.11**	**0.98**	**0.913**
	XGBoost (Random Search)	**0.06**	**0.04**	**1.00**	**0.923**
	Random Forest	**0.04**	**0.02**	**1.00**	**0.940**
1095 Days	TFT	**0.25**	**0.15**	**0.95**	**0.969**
	XGBoost	**0.06**	**0.05**	**1.00**	**0.918**
	XGBoost (Random Search)	**0.12**	**0.08**	**0.99**	**0.916**
	Random Forest	**0.08**	**0.06**	**0.99**	**0.966**

Table 2
Monthly data.

Prediction Horizon	Model Type	Train			Holdout NSE
		RMSE	MAE	R^2	
12 months	TFT	**0.35**	**0.27**	**0.85**	**−1.276**
	XGBoost	**0.04**	**0.03**	**1.00**	**0.747**
	XGBoost (Random Search)	**0.23**	**0.13**	**0.97**	**−5.814**
	Random Forest	**0.17**	**0.08**	**0.89**	**−2.362**
24 months	TFT	**0.90**	**0.56**	**0.13**	**0.162**
	XGBoost	**0.02**	**0.01**	**1.00**	**0.862**
	XGBoost (Random Search)	**0.25**	**0.13**	**0.96**	**−5.070**
	Random Forest	**0.45**	**0.25**	**0.37**	**−0.159**
30 months	TFT	**0.85**	**0.51**	**0.56**	**−40.943**
	XGBoost	**0.02**	**0.01**	**1.00**	**−5.157**
	XGBoost (Random Search)	**0.18**	**0.11**	**0.95**	**−0.295**
	Random Forest	**0.59**	**0.33**	**0.39**	**−15.383**

patterns. However, forecasting long-term trends can be difficult due to the accumulation of prediction horizons; small errors might accumulate, resulting in higher RMSE and lower NSE values (Wu et al., 2020a). Furthermore, long-term forecasting is more susceptible to unexpected external factors such as major weather occurrences, land use changes, and policy changes. These variables provide additional unpredictability and uncertainty, which the model may not account for, resulting in poor performance (Deng et al., 2023). In addition to the correlations between NH_4^+ concentration and environmental factors may shift over time as a result of seasonal climate changes and human activities (Tang et al., 2022). Models such as XGBoost and TFT may struggle to retain accuracy over longer timescales if they are unable to adapt to these changing dynamics (Wulfmeyer et al., 2011; Guo et al., 2021).

Whilst the models worked well on the River Lee dataset, its capacity to generalise to other rivers or geographic areas may be limited by the hydrological and environmental conditions unique to the River Lee. Rivers in various geographic areas may have varied water quality dynamics due to variances in temperature, land use, pollution sources, and hydrological patterns (Egbueri et al., 2023). Thus, the models' generalisability to other areas would most likely be determined by how comparable those ecosystems were to the circumstances found in the River Lee. To make the model more applicable to different rivers, it might be retrained on a regular basis with local data from other geographic regions. Furthermore, including exogenous factors, such as climate projections or land-use data, may improve its capacity to generalise across varied situations. The holdout set performance implies that the model can make reliable predictions in similar ecosystems, but more testing on more river systems is needed to demonstrate its broader application.

3.3. Detailed analysis of findings

It is important to note that Table 1 and Table 2 represent the training results, which show how well the models match the training data. On the other hand, Fig. 2 and Fig. 7 illustrate the holdout RMSE of the validation results, showing the model's performance on new, unseen data. The differentiation is significant because it emphasises the model's capacity to generalise outside of the training, which is essential for accurate long-term forecasts.

Moreover, the importance of environmental factors in forecasting NH4 levels was determined using feature importance metrics supplied by machine learning algorithms, particularly XGBoost and Random Forest. These models rank attributes depending on how they help to reduce prediction error. Dissolved oxygen (O_2) appeared as one of the most important predictors, suggesting its vital role in regulating ammonium concentrations via processes like as nitrification and oxygen depletion (Qiao et al., 2020). Temperature and pH were also consistently important, altering the balance of ammonium and unionised ammonia, which varies with temperature and water chemistry.

The significance of these factors varies throughout time spans. Variables like temperature and dissolved oxygen have a greater influence on short-term forecasts (daily) since they fluctuate rapidly in response to daily environmental circumstances. Over longer time scales (e.g., monthly or multi-year predictions), parameters like turbidity and conductivity become more important since they represent wider hydrological and biogeochemical processes, such as seasonal runoff patterns and nitrogen cycling. This shift in variable importance emphasises the intricate interaction of physical, chemical, and biological components in river ecosystems, with short-term projections more susceptible to rapid environmental changes and long-term forecasts impacted by bigger, slower-moving processes.

3.3.1. Predicting 3 years based on daily data

The findings highlight how well ML models predict the ammonium levels in river ecosystems. The robust performance of the models over both long- and short-term horizons represents a substantial development in environmental modelling, and essential tool for proactive water management. Fig. 2 represent the RMSE of the holdout set which shows how the models are performing on new unseen data.

The capacity to anticipate NH_4^+ levels up to 3 years in the future using daily data is a noticeable achievement in machine learning. Daily data offer a high-resolution temporal framework that captures the complex fluctuations in environmental circumstances required to describe the dynamics of ammonium levels effectively. The Random Forest Regressor (RF) model performed well in this long-term forecasting, as shown in Fig. 2. This model's resilience is demonetised by its capacity to sustain high R^2 (0.97), and low RMSE of (0.18) over 3 years, indicating that it successfully represents the complex interactions between ammonium and the other environmental variables such as temperature, turbidity, chlorophyll, dissolved oxygen, conductivity, and pH. This model's robustness is especially notable since it continuously obtains low RMSE values across several steps, indicating steady and reliable performance.

To test the model's robustness, we illustrated and analysed model performance across a 1095-day prediction horizon. The figures below

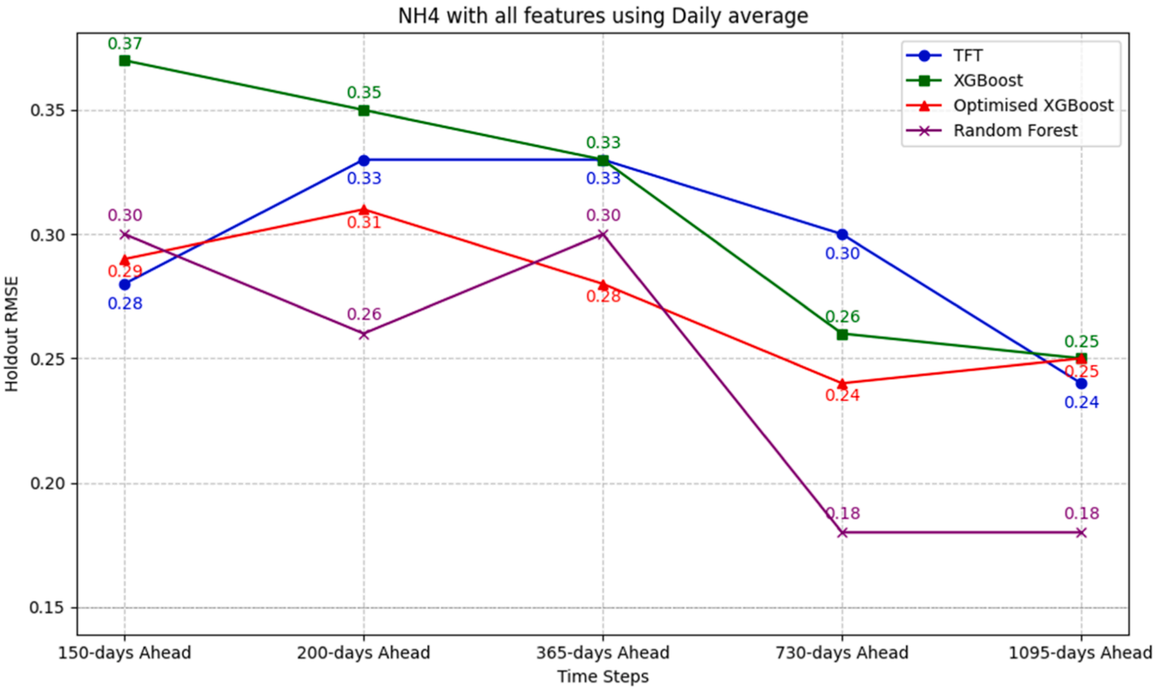

Fig. 2. The holdout RMSE in the daily prediction.

demonstrate the actual versus the predicted ammonium levels over a period of 1095 days, highlighting the model's ability to track observed data patterns closely. Furthermore, the distribution of prediction errors, which usually shows a tight clustering around zero, indicates good prediction accuracy and dependability (Pinson and Kariniotakis, 2004).

Fig. 3 illustrates the performance of RF in predicting the 1095 days; Fig. 3a shows a good fit to the actual ammonium levels, closely matching the observed data patterns. This suggests that the models accurately represent the underlying patterns and fluctuations in the data (Khozani et al., 2022). Furthermore, Fig. 3b shows a small, narrow and centred error distribution around 0, indicating great precision and low bias in the model's predictions. This shows how well the RF handles long-term acting.

The model is quite narrow and centred, with the majority of errors clustered between −0.2 and 0.2. This suggests that the model is typically pretty accurate in its predictions over the 1095-day timeframe. The distribution is somewhat tilted to the right, with a few errors larger than 0.6, but these are outliers. This behaviour indicates that Random Forest operates well under typical settings but may fail to manage extreme values or rapid changes in environmental parameters. The model's stability, as evidenced by the error distribution, suggests that it is well suited for long-term forecasts in stable contexts. Its capacity to give consistent predictions is due to its ensemble learning method, which combines the outcomes of several decision trees (Breiman, 2001). This strategy reduces overfitting and improves model resilience, especially in datasets with noise and unpredictability. However, the occurrence of a few high-error outliers indicates that the model may fail to capture all complicated interactions between environmental factors in extreme circumstances. Future enhancements might include adding more environmental features or adjusting the model to better handle outlier events.

The TFT model captures overall trends and seasonal patterns, but it struggles with high values and has a bigger error distribution. It has a larger error distribution than Random Forest, ranging from −0.75–0.75. While the distribution's apex stays centred around 0, the longer tails on both sides show that the TFT model is more prone to higher prediction mistakes over longer time horizons. This suggests that, while the model is effective for broad forecasts, it may not be dependable for accurately anticipating exceptional events. This drawback might be attributed to the TFT model's sensitivity to data granularity since short-term swings are efficiently recorded, but longer-term patterns result in compounding inaccuracies. The broader error distribution, as shown in Fig. 4b, indicates that the model may not be capturing all underlying processes or that it is more prone to overfitting on shorter patterns, which becomes less valid over longer time horizons. The errors are more uniformly distributed, with a clear tail towards higher error values, showing that certain forecasts vary more considerably from the actual values.

Many approaches may be considered to deal with these challenges. To minimise compounding errors over time, the model may benefit from aggregating data (Zeng et al., 2024) into wider periods, such as weekly, to focus on general trends rather than short-term fluctuations.

The XGBoost model is effective at accurately monitoring real ammonium levels while preserving a tight error distribution. It is more concentrated than the TFT model, although it has a tiny right skew, with most errors falling between −0.2 and 0.4. The distribution's peak is somewhat moved to the right of 0, showing that the model tends to overestimate ammonium levels. While this improves overall accuracy and reduces significant mistakes, it may add a small bias in predictions, especially in datasets with unbalanced distributions or skewed variables. Calibration or regularisation strategies might be investigated to solve the persistent over-prediction and enhance the model's performance. Nonetheless, the model's capacity to avoid huge mistakes makes it an excellent option for long-term predictions since it stays consistent even when dealing with complicated interactions between environmental factors., however across all the models, this model had the highest RMSE, as shown in Fig. 2. XGBoost's good performance can be due to its adept handling of non-linear correlations and interactions between features, as Wang et al. (2021) presented in their study. The model's accuracy in long-term forecasting demonstrates its applicability for applications that need a thorough grasp of temporal dynamics, such as anticipating seasonal variations and responding to slow adjustments in environmental circumstances

Fig. 6 illustrates the performance of model XGBoost with optimiser across 1095 days; this model is the second-best model after RF. With closely matched actual and predicted values, as seen in Fig. 6a. The optimisation process enhanced the accuracy of this model (Sun, 2020) to

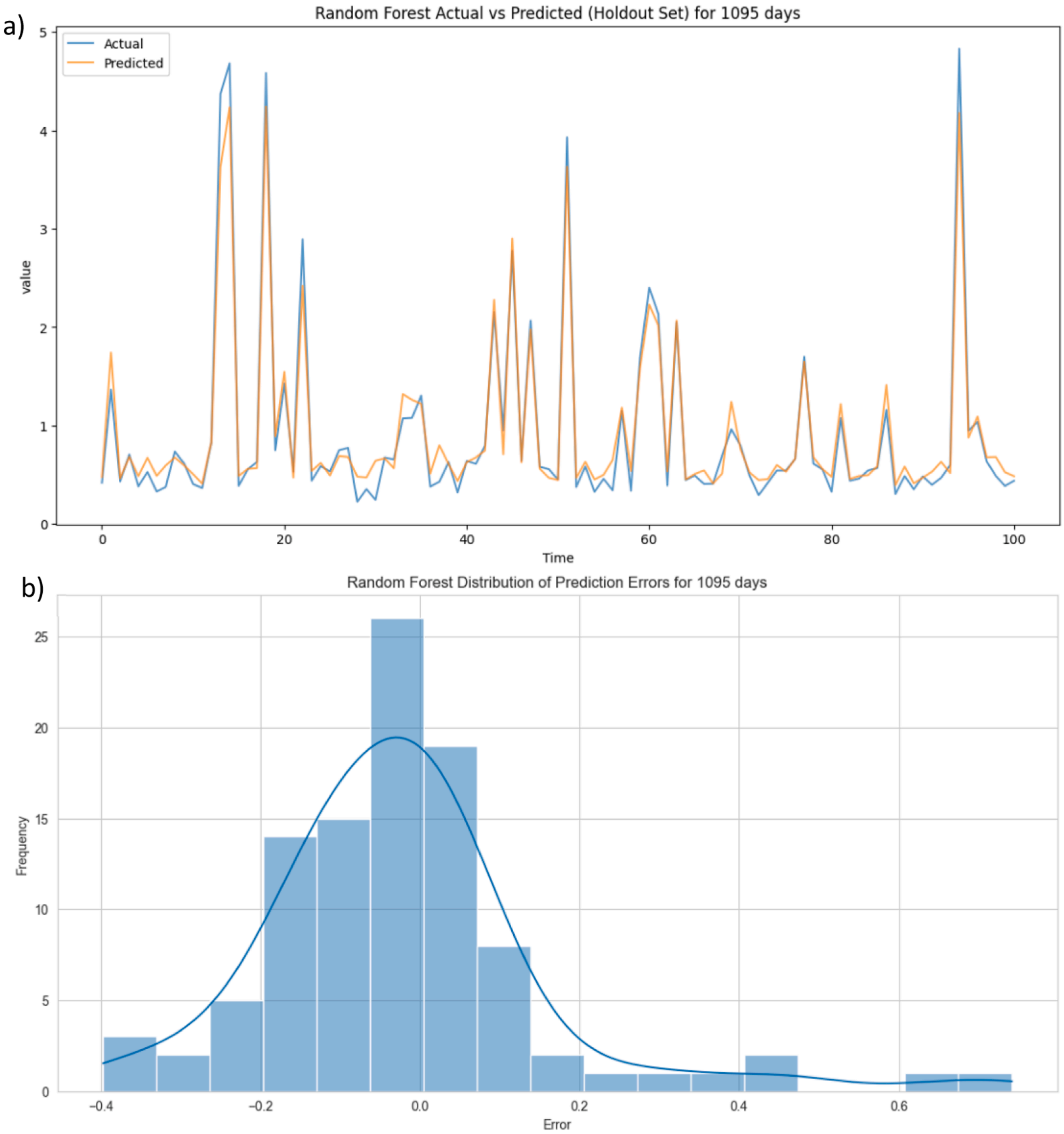

Fig. 3. a) RF actual vs predicted across 1095 days b) RF distribution of errors for 1095 days.

predict ammonium levels over extended horizons. The model's capacity to generalise from the training data to unknown data was probably strengthened by the use of random search optimisation, which adjusted the hyperparameters that control the model's learning process. This version has been optimised to better handle complicated relationships and better adjust to the inherent unpredictability in the environment.

From the hydrological perspective, estimating ammonium levels with such accuracy over long periods is essential. Long-term forecasting allows water resource managers to foresee and minimise prospective ammonium toxicity concerns, such as oxygen depletion and the stress caused to aquatic life (Jia et al., 2023). Predicting changes in water quality over a 3-year period allows for proactive interventions to reduce nutrient loads, regulate water flows, and maintain ideal environmental conditions. This predictive capacity facilitates the creation of long-term water management systems that can adapt to changing climate and human effects.

Figs. 3–6 illustrate that the TFT model has a larger error distribution, especially for longer prediction horizons. This might be due to a variety of circumstances. First, data granularity is important, since the model covers short-term variations well but suffers with long-term forecasts when tiny mistakes compound. Over time, these compounding mistakes produce larger error distributions, as shown in multi-horizon forecasting. Furthermore, the TFT model is susceptible to abrupt changes or outliers in environmental variables, such as temperature or oxygen levels. The attention mechanism in TFT, which is useful for short-term prediction, may become less efficient if the associations between variables change slowly or unexpectedly over time, resulting in decreased performance. Certain characteristics, including as turbidity and chlorophyll, may also contribute to increasing inaccuracy since their dynamics are impacted by complex and nonlinear environmental interactions that the model may not completely reflect in long-term projections.

This error study indicates that, whilst the TFT model excels at capturing immediate trends, modifications are required for longer-term forecasting. Future research might look into data aggregation approaches or model regularisation procedures to improve the model's capacity to generalise across longer time periods. Furthermore, including more exogenous variables, such as climate or land-use data,

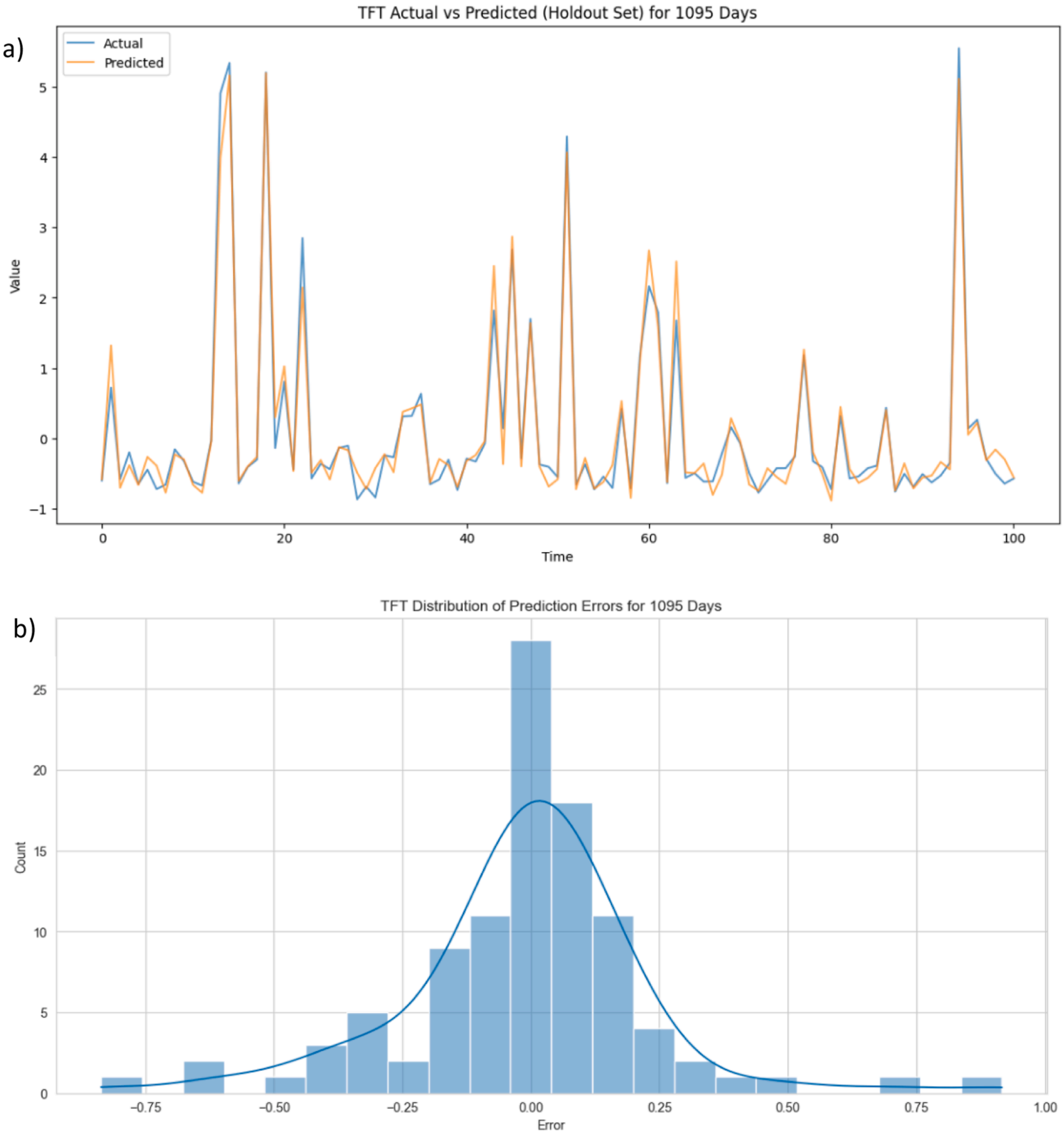

Fig. 4. a) TFT actual vs predicted across 1095 days b) TFT distribution of errors for 1095.

may allow the model to better account for external influences impacting the overall error distribution.

3.3.2. *Predicting 30 months based on monthly data*

Fig. 7 presents the holdout RMSE for estimating the ammonium levels based on monthly data at various prediction horizons (12, 24, and 30) months ahead. The finding demonstrates the efficacy of the XGBoost model, which continually maintained a low RMSE across all horizons. The model's performance demonstrates its capacity to capture long-term trends and seasonal fluctuations in ammonium levels, making it an important tool for hydrological forecasting.

The capacity of the XGBoost model to generalise successfully from monthly aggregated data is evidenced by its consistent RMSE values, which demonstrate its robustness in dealing with long-term predictions. This is especially essential for strategic planning and policy making, as understanding seasonal peaks and long-term patterns in the ammonium level may help direct the timing of measures like improved monitoring or nutrient reduction programmes to reduce the risk of ecological damage.

The TFT model, on the other hand, showed rising RMSE values over larger time frames, indicating difficulties in sustaining accuracy over protracted durations. This might be owing to the model's complexity and the possibility of overfitting to shorter-term trends, which was a main focus in Ali et al., (2024)'s study, resulting in worse performance over longer time periods. The Random Forest model demonstrated consistent performance with generally steady RMSE values, showing its capacity to handle monthly aggregated data successfully.

Precise long-term forecasting of ammonium levels can effectively prevent such ecological disruption. Ammonium prediction can help in proactive steps to minimise nutrient inputs (Li et al., 2014), including restricting agricultural runoff or improving wastewater treatment facilities (Kube et al., 2018). Additionally, by reducing the amount of ammonium entering river systems, these actions can stop the circumstances that cause hypoxia (Geeraert et al., 2021). Where a body of water's oxygen concentration drops below the required levels to sustain the majority of marine life. Specifically, dissolved oxygen values of less than 2 mg/l are frequently used to characterise hypoxic environment. Long term hypoxia can lower biodiversity since only a small number of

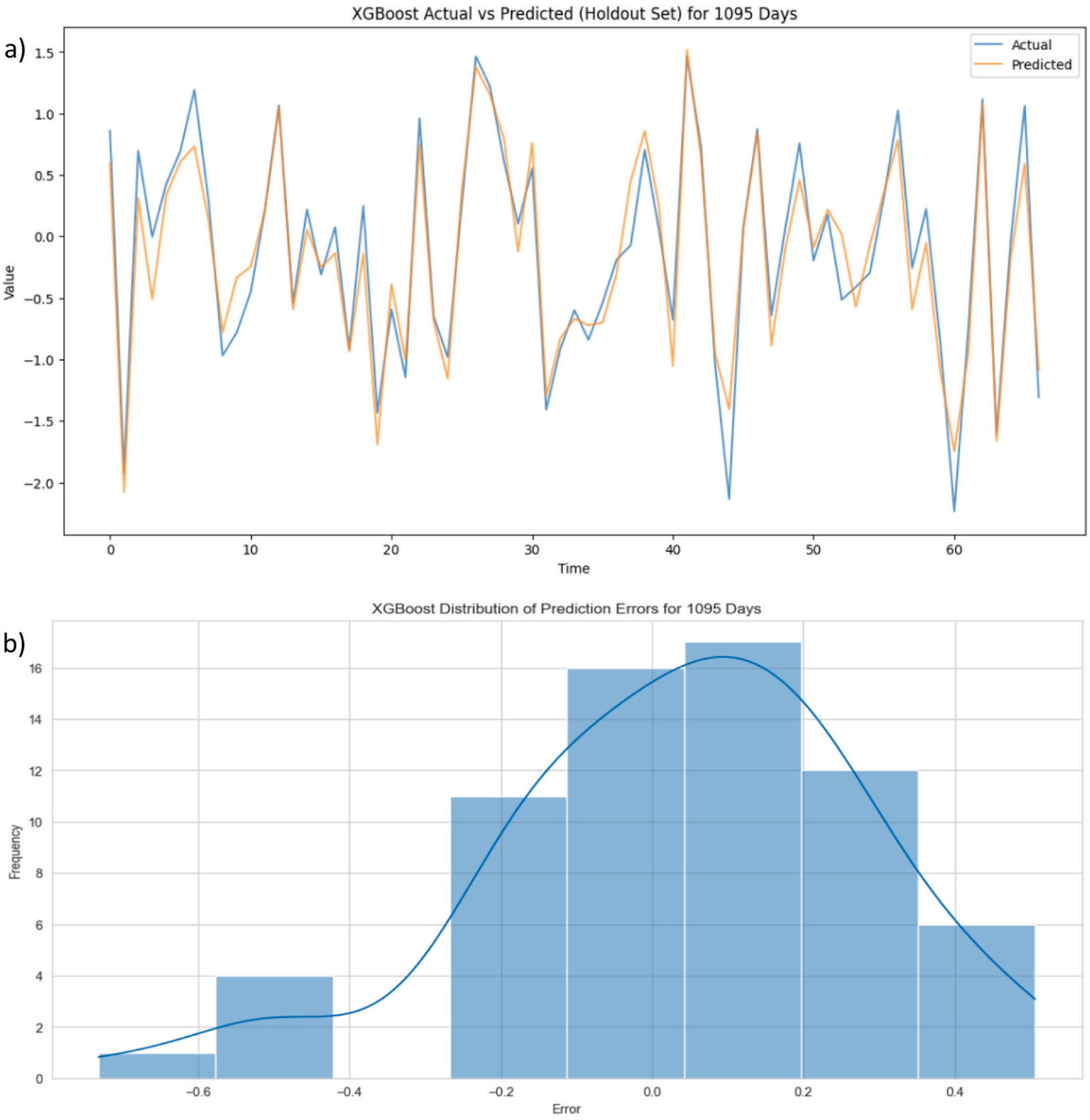

Fig. 5. a) XGBoost actual vs predicted across 1095 days b) XGBoost distribution of errors for 1095.

tolerant species can endure. Furthermore, the insight gained from the models can be a tool to provide information about the timing and the strength of treatment. Monitoring the seasonal and long-term pattern in NH_4^+ concentration help in planning the best times to take measures like aerating water, flushing out water bodies or using chemical treatments to neutralize ammonium.

This study emphasises the complex interaction between ammonium levels and numerous environmental conditions. The mild negative connection between NH_4^+ and O_2 shown in Fig. 1, as well as the positive correlation with temperature, aligns with accepted hydrological concepts that emphasise the impact of oxidation processes (Liu and Wang, 2023) and thermal dynamics (Pei et al., 2015) on water quality. These linkages are critical for understanding how changes in environmental circumstances caused by variables such as climate change and human activity affect river ecosystems.

The findings also emphasise the importance of integrated water management techniques that account for the intricate interplay between physical, chemical and biological processes in river systems. For example, maintaining normal flow regimes and temperature conditions is critical for sustaining dissolved oxygen levels (Dowling and Wiley, 1986). Furthermore, the base flow and storm flow of rivers have a major influence on the ammonium concentration. Storms and other high flow events can cause more runoff from urban and agricultural regions, which can increase the amount of ammonium that enters rivers (Lin et al., 2022). Ammonium levels are also influenced by sediment transport processes since ammonium can be absorbed by sediments and released back into the water column when sediment transport dynamics are altered, as they can be during high flow episodes.

Machine learning models give practical insight into water quality management. Short-term forecasting using TFT model can assist water authorities in implementing re-time interventions, such as altering oxygenation levels or treatments to avoid hypoxia (Geeraert et al., 2021). Long-term predictions for XGBoost and RF allow for strategic planning of infrastructure expenditures, such as updating wastewater treatment plants in response to certain environmental circumstances, such as changing water flow during periods of low dissolved oxygen (Dowling and Wiley, 1986).

Nitrification and denitrification are two important nitrogen cycle that control ammonium levels in rivers (Qiao et al., 2020). Under aerobic settings, nitrification transforms ammonium into nitrate, whereas

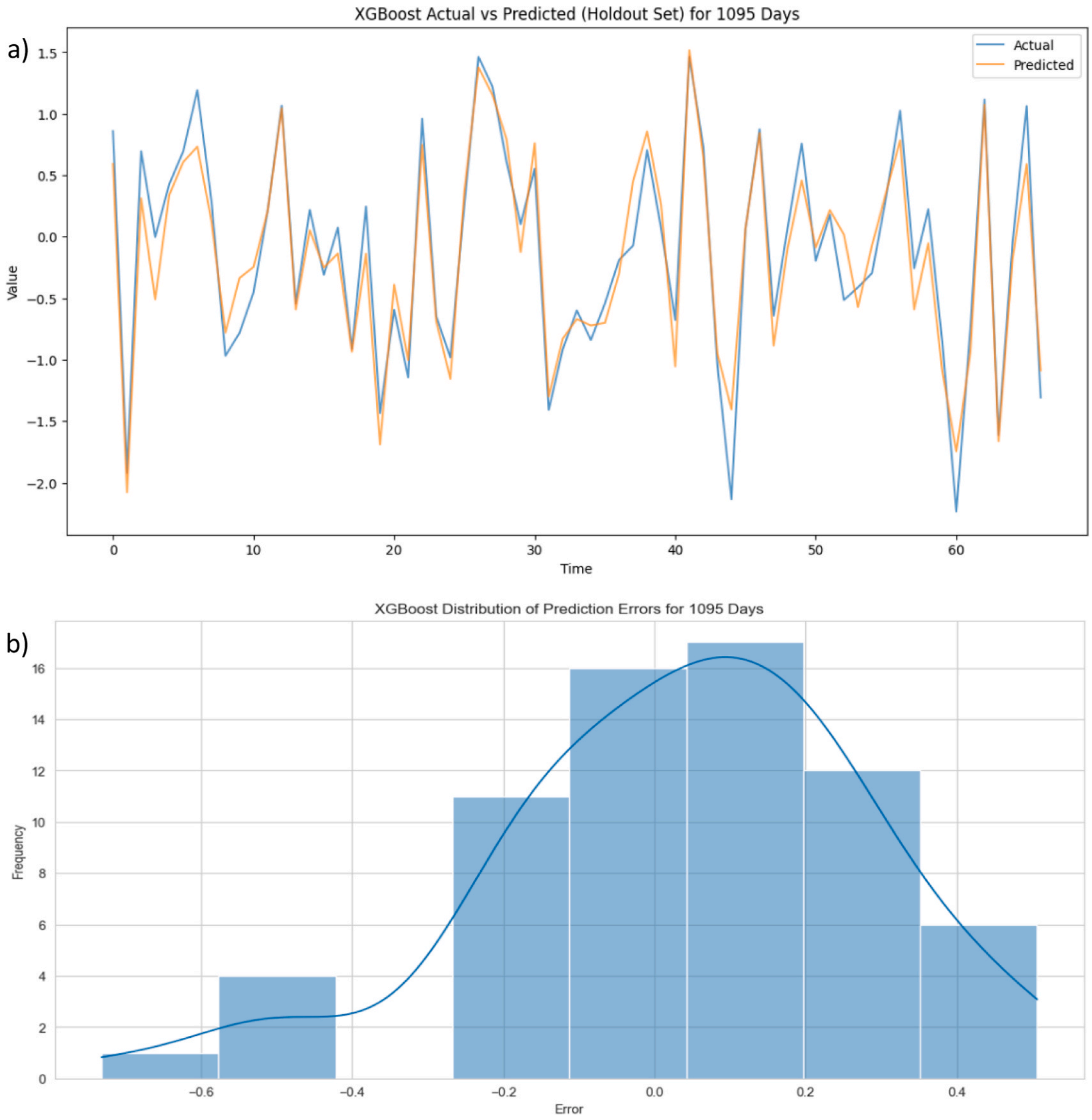

Fig. 6. a) XGBoost with optimiser actual vs predicted across 1095 days b) XGBoost with optimiser distribution of errors for 1095.

under anaerobic conditions, denitrification lowers nitrate to nitrogen gas. These processes are influenced by variables like pH, temperature and dissolved oxygen concentrations. In sediments, ammonium may be absorbed into organic materials and clay particles (Fan et al., 2021). The process of absorption-desorption can be affected by variations in ambient factors like pH and iconic strength impacting ammonium concentration in water.

Ammonium is necessary for the growth of aquatic plants and algae (Akinnawo, 2023), and elevated ammonium levels can cause algal blooms. When these blooms decompose, the amounts of dissolved oxygen in water will decrease, leading to exacerbating hypoxia. Additionally, microorganisms are essential to the nitrogen cycle because they mediate nitrification and denitrification. The activity of these microbes is affected by variables such as temperature, organic carbon availability and oxygen levels, which together work to control the concentration of ammonium in the water (Pajares and Ramos, 2019).

The combination of machine learning algorithms and extensive environmental information creates a potential way to forecast and regulate water quality. The better effectiveness of models such as XGBoost with random search optimisation indicates how modern data-driven approaches may successfully capture the complex dynamics influencing ammonium levels. This integrated method enables more accurate and trustworthy projections, which are essential for establishing proactive water management policies.

Since TFT has multi-time horizon processing and attention-based architecture, it is recognised for being computationally demanding system, which may limit its application in real-time or resource-constrained environments (Wang et al., 2024). In comparison to the RF and XGBoost models, this makes the TFT model more resource-demanding in terms of memory and processing time. Nevertheless, the model's capacity to identify intricate temporal patterns and provide incredible precise short-term forecasts more than makes up for the increase computing expense. Moreover, the models assume that the correlations between environmental factors and ammonium levels are constant over time; however, disruptions caused by climate change, land-use changes, or human activities may impair the accuracy of long-term projections.

On the other hand, RF and XGBoost models are more effective for long-term forecasting since they use less computing power and can handle big datasets with missing values as shown in Table 1 and Table 2. Whilst TFT excels in short-term prediction accuracy, in longer horizons XGBoost and RF provided a superior balance between computational efficiency and predictive effectiveness. To further increase the model's efficiency, the XGBoost Random Search optimiser enables efficient tuning without using a lot of resources.

Although the models used in this study was trained using historical

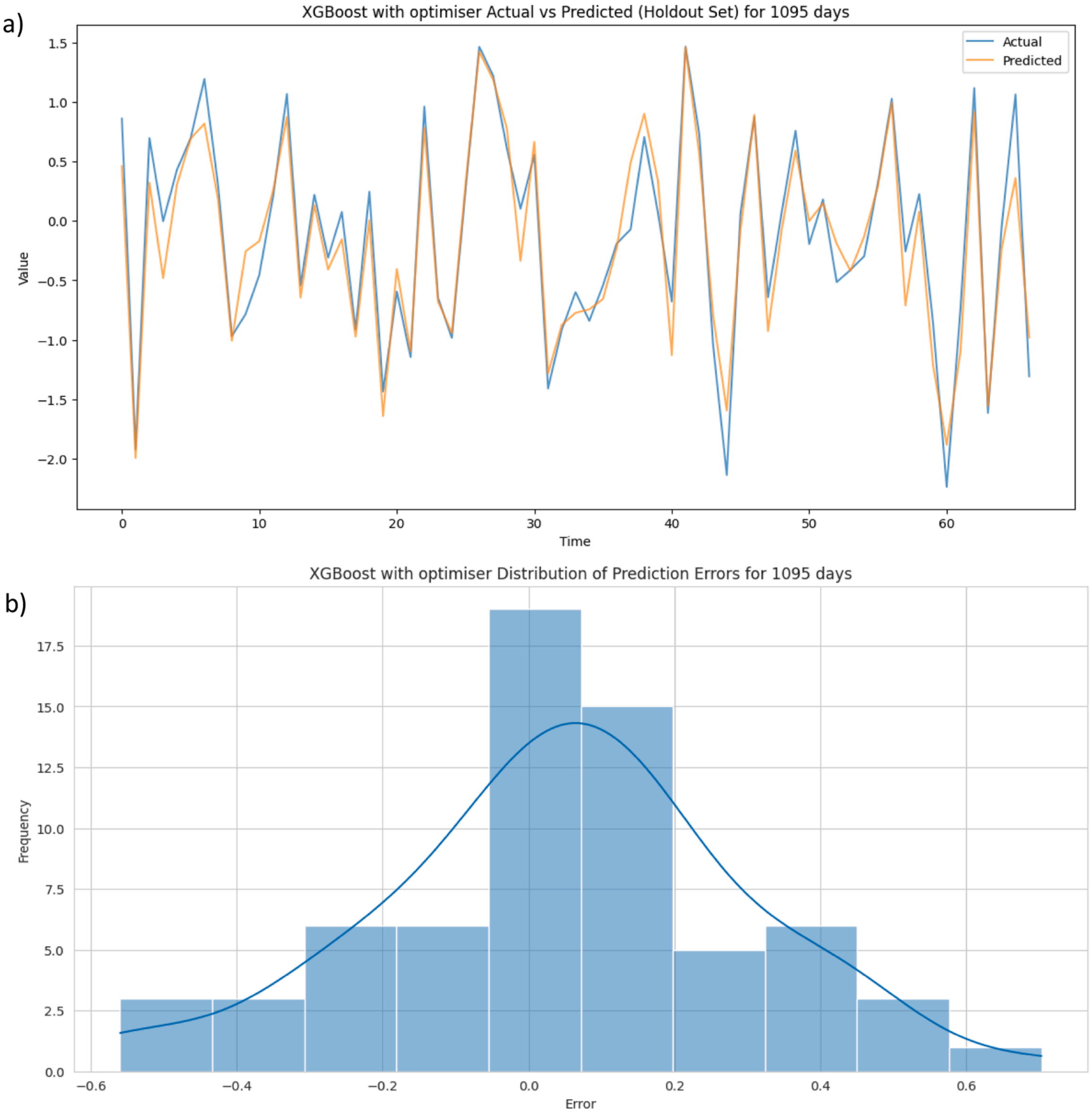

Fig. 7. The holdout RMSE in the monthly prediction.

data to forecast ammonium levels it is difficult to account for any future changes in the environment or human activity that might not have been captured in the training set. Climate change land-use shifts, urbanisation, and changes in agricultural techniques might all have a large impact on ammonium dynamics that historical patterns do not account for. This is a challenge for long-term forecasts, as the model's capacity to generalise may deteriorate if situations vary from prior trends. To address this, future iterations of the program might include scenario-based forecasting techniques (Li et al., 2018). The models might be modified to mimic probable fur environmental alterations by including external information such as climate predictions, land-use models, or forecasts of policy change. Furthermore, periodic retraining with updated data might help keep the model relevant as new information becomes available (Angel and McCabe, 2022). This would enable the mode to adapt to new trends in environmental factors and human activities, hence enhancing its long-term forecast accuracy.

The hydrological insights gained from these models are critical for understanding and regulating the processes that determine ammonium levels in river ecosystems. These models help to establish effective management plans for mitigating the effects of environmental changes and human activities on water quality by accurately anticipating future circumstances. This integrated strategy improves our capacity to anticipate and manage the health of aquatic ecosystems, preserving their long-term viability in the face of persistent environmental threats.

For water management authorities, the contributions of this study offer insightful information, especially when it comes to forecasting ammonium levels over the short and long term. Water resource managers can take pre-emptive steps to reduce ammonium concentrations in rivers since models like Random Forest and XGBoost can produce reliable long-term forecasts (up to three years). For instance, anticipating rises in ammonium levels can assist authorities in organising corrective measures like streamlining wastewater treatment procedures, cutting down on agricultural runoff, or putting in place regulations to manage fertiliser inputs. These predictions can inform policy adjustments, such as implementing stricter fertiliser application limits during seasons of high runoff to avoid excess ammonium from entering rivers. In urban settings, authorities might update wastewater discharge rules depending on expected ammonium levels, ensuring that treatment plants adjust operations during high-risk times. Proactive policies based on realistic model projections can result in more sustainable water management practices, decreasing the ecological effect of human activities on river systems.

Real-time monitoring is possible with short-term forecasts, which is where the TFT excels. These can direct quick hus, including raising the water's oxygen content to stop hypoxia when ammonium levels rise. The models' ability to provide both short- and long-term projections can help in the development of more flexible and successful water management plans, enhancing the general health of rivers and lowering ecological threats like oxygen depletion and toxic algal blooms.

4. Conclusion

This paper advanced the use of machine learning in environmental science, especially for forecasting ammonium NH_4^+ levels in river ecosystems. We created a solid framework by combining hydrological knowledge with powerful machine learning algorithms, improving our capacity to anticipate water quality in the short and long term. The key finding of this study are:

- This is the first research to predict ammonium levels up to 3 years in advance using daily data and up to 30 months ahead using monthly data. This dual-scale forecasting capacity offers flexibility in addressing a variety of environmental concerns and is essential for both short- and long-term water management planning. This is a paradigm change in how machine learning may be applied to ecological forecasting, bringing new levels of interpretability and precision.
- Our technique, which includes a dynamic evaluation of several environmental parameters such as temperature, pH, turbidity, chlorophyll, dissolved oxygen, and conductivity. This thorough evaluation gives a complete knowledge of the circumstances that might contribute to ammonia toxicity in river ecosystems.
- The created framework is scalable to other river systems than the United Kingdom, making it an important tool for worldwide water sustainability programmes. This scalability means that the model may be utilised in a variety of environmental scenarios, increasing its usefulness and effect.
- This study's findings promote the development of proactive water management measures. Water resource managers can employ precise forecasting to avoid possible concerns such as oxygen depletion and biological stress caused by increased ammonium levels.
- We also investigated that precise forecasts enable prompt nutrient load reduction initiatives, including using best management practices (BMPs) in agriculture to regulate fertiliser application and runoff.

First, the use of advanced machine learning models, namely Temporal Fusion Transformer (TFT), for multi-horizon ammonium level forecasting is a noteworthy breakthrough. The capacity of the TFT model to deal with both short-term fluctuation and long-term trends establishes a new standard for forecast accuracy in complicated river systems such as the River Lee, where previous models frequently fail.

Third, by making long-term forecasting, this study fills a gap in ammonium level prediction, particularly in respect to multi-scale environmental variables such as seasonal climate change and human activity. The combination of short and long-term perspectives distinguishes this study, which provides a flexible and comprehensive prediction framework that can be immediately applied to the policymakers. Finally, the study's multi-model strategy, which incorporates Random Forest, XGBoost, and TFT, gives compelling evidence of how diverse machine learning approaches may work together to improve overall knowledge of environmental processes. This study sets a new benchmark for future research in the predictive modelling of water quality and other environmental issues.

Author contributions

A.J. is primarily responsible for carrying out the research, developing the models and writing the initial drafts of the manuscript. A.J. performed the data analysis, created the figures and drafted the original manuscript. A.A. provided direction and supervision during the writing and research phases. He provided important ideas that influenced the research's direction and contributed to the study's conception and design. IA.A. also helped with editing and rewriting the final document.

CRediT authorship contribution statement

Ashraf Ahmed: Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Ali Ali**: Writing – review & editing, Writing – original draft, Validation, Software, Methodology, Investigation, Formal analysis, Data curation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

We would like to thank the Environment Agency of England for providing the data. The authors would like to thank the anonymous reviewers for their constructive comments that helped improve this manuscript. The study was partially funded by UKRI project 10063665.

Data availability

Data will be made available on request.

References

Abou Omar, K.B., 2018. XGBoost and LGBM for Porto Seguro's Kaggle challenge: A comparison. Preprint Semester Project.

Ahmed, A.A., Sayed, S., Abdoulhalik, A., Moutari, S., Oyedele, L., 2024. Applications of machine learning to water resources management: a review of present status and future opportunities. J. Clean. Prod., 140715

Akinnawo, S.O., 2023. Eutrophication: causes, consequences, physical, chemical and biological techniques for mitigation strategies. Environ. Chall., 100733

Ali, A.J., Ahmed, A.A., and Abbod M.F., 2024. Groundwater level predictions in the Thames Basin, London over extended horizons using Transformers and advanced machine learning models. Journal of cleaner production, Under Review.

Amor, L.B., Lahyani, I., Jmaiel, M., 2016. Recursive and rolling windows for medical time series forecasting: a comparative study (August). IEEE, pp. 106–113 (August).

Angel, Y., McCabe, M.F., 2022. Machine learning strategies for the retrieval of leaf-chlorophyll dynamics: model choice, sequential versus retraining learning, and hyperspectral predictors. Front. Plant Sci. 13, 722442.

Appels, W.M., Graham, C.B., Freer, J.E., McDonnell, J.J., 2015. Factors affecting the spatial pattern of bedrock groundwater recharge at the hillslope scale. Hydrol. Process. 29 (21), 4594–4610.

Ayejoto, D.A., Egbueri, J.C., Enyigwe, M.T., Chiaghanam, O.I., Ameh, P.D., 2022. Application of HMTL and novel IWQI models in rural groundwater quality assessment: a case study in Nigeria. Toxin Rev. 41 (3), 918–932.

Azrour, M., Mabrouki, J., Fattah, G., Guezzaz, A., Aziz, F., 2022. Machine learning algorithms for efficient water quality prediction. Model. Earth Syst. Environ. 8 (2), 2793–2801.

Bergstra, J., Bengio, Y., 2012. Random search for hyper-parameter optimization. J. Mach. Learn. Res. 13 (2).

Bhatnagar, A., Sillanpää, M., 2011. A review of emerging adsorbents for nitrate removal from water. Chem. Eng. J. 168 (2), 493–504.

Biau, G., Scornet, E., 2016. A random forest guided tour. Test 25, 197–227.

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

Britto, D.T., Kronzucker, H.J., 2002. NH4+ toxicity in higher plants: a critical review. J. Plant Physiol. 159 (6), 567–584.

Cerqueira, V., Torgo, L., Mozetič, I., 2020. Evaluating time series forecasting models: an empirical study on performance estimation methods. Mach. Learn. 109, 1997–2028.

Chapman, D., 1996. A guide to use of biota, sediments and water in environmental monitoring. Water quality assessment. 19972nd ed. London and New York UNESCO, WHO, UNEP, 626.

Chen, T. and Guestrin, C., (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Chicco, D., Warrens, M.J., Jurman, G., 2021. The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. PeerJ Comput. Sci. 7, e623.

Cirillo, D., Núñez-Carpintero, I., Valencia, A., 2021. Artificial intelligence in cancer research: learning at different levels of data granularity. Mol. Oncol. 15 (4), 817–829.

Clevert, D.A., Unterthiner, T., Hochreiter, S., 2015. Fast and accurate deep network learning by exponential linear units (elus). arXiv Prepr. arXiv 1511, 07289.

Covatti, G., Grischek, T., 2021. Sources and behavior of ammonium during riverbank filtration. Water Res. 191, 116788.

Dauphin, Y.N., Fan, A., Auli, M., Grangier, D., 2017. July. Language modeling with gated convolutional networks. Int. Conf. Mach. Learn. 933–941.

De Vet, W.W.J.M., Van Genuchten, C.C.A., Van Loosdrecht, M.C.M., Van Dijk, J.C., 2010. Water quality and treatment of river bank filtrate. Drink. Water Eng. Sci. 3 (1), 79–90.

Deng, Y., Ye, X., Du, X., 2023. Predictive modeling and analysis of key drivers of groundwater nitrate pollution based on machine learning. J. Hydrol. 624, 129934.

Dingman, S.L., 2015. Physical hydrology. Waveland press.

Dowling, D.C. and Wiley, M.J., 1986. The effects of dissolved oxygen, temperature, and low stream flow on fishes: a literature review. Illinois Natural History Survey Technical Reports.

Duc, L., Sawada, Y., 2023. A signal-processing-based interpretation of the Nash–Sutcliffe efficiency. Hydrol. Earth Syst. Sci. 27 (9), 1827–1839.

Dugdale, S.J., Malcolm, I.A., Hannah, D.M., 2024. Understanding the effects of spatially variable riparian tree planting strategies to target water temperature reductions in rivers. J. Hydrol. 635, 131163.

Dwork, C., Feldman, V., Hardt, M., Pitassi, T., Reingold, O., Roth, A., 2015. Generalisation in adaptive data analysis and holdout reuse. Adv. Neural Inf. Process. Syst. 28.

Egbueri, J.C., Unigwe, C.O., Omeka, M.E., Ayejoto, D.A., 2023. Urban groundwater quality assessment using pollution indicators and multivariate statistical tools: a case study in southeast Nigeria. Int. J. Environ. Anal. Chem. 103 (14), 3324–3350.

Environment Agency, Water quality monitoring (2014).-WaterqualitydatainterpretationfornontechnicalcustomersFeb201.pdf (moderngov.co.uk).

Fan, X., Xue, Q., Liu, S., Tang, J., Qiao, J., Huang, Y., Sun, J., Liu, N., 2021. The influence of soil particle size distribution and clay minerals on ammonium nitrogen in weathered crust elution-deposited rare earth tailing. Ecotoxicol. Environ. Saf. 208, 111663.

Fayer, G., Lima, L., Miranda, F., Santos, J., Campos, R., Bignoto, V., Andrade, M., Moraes, M., Ribeiro, C., Capriles, P., Goliatt, L., 2023. A temporal fusion transformer deep learning model for long-term streamflow forecasting: a case study in the funil reservoir, Southeast Brazil. Knowl. -Based Eng. Sci. 4 (2), 73–88.

Geeraert, N., Archana, A., Xu, M.N., Kao, S.J., Baker, D.M., Thibodeau, B., 2021. Investigating the link between Pearl River-induced eutrophication and hypoxia in Hong Kong shallow coastal waters. Sci. Total Environ. 772, 145007.

Gislason, P.O., Benediktsson, J.A., Sveinsson, J.R., 2006. Random forests for land cover classification. Pattern Recognit. Lett. 27 (4), 294–300.

Goldstein, B.A., Polley, E.C., Briggs, F.B., 2011. Random forests for genetic association studies. Stat. Appl. Genet. Mol. Biol. 10 (1).

González-Enrique, J., Ruiz-Aguilar, J.J., Madrid Navarro, E., Martínez Álvarez-Castellanos, R., Felis Enguix, I., Jerez, J.M. and Turias, I.J., 2022, September. Deep Learning Approach for the Prediction of the Concentration of Chlorophyll ɑ in Seawater. A Case Study in El Mar Menor (Spain). In International Workshop on Soft Computing Models in Industrial and Environmental Applications (pp. 72-85). Cham: Springer Nature Switzerland.

Groeschke, M., Frommen, T., Winkler, A., Schneider, M., 2017. Sewage-borne ammonium at a river bank filtration site in central Delhi, India: simplified flow and reactive transport modeling to support decision-making about water management strategies. Geosciences 7 (3), 48.

Guo, J., Zhang, J., Yang, K., Liao, H., Zhang, S., Huang, K., Lv, Y., Shao, J., Yu, T., Tong, B., Li, J., 2021. Investigation of near-global daytime boundary layer height using high-resolution radiosondes: first results and comparison with ERA5, MERRA-2, JRA-55, and NCEP-2 reanalyses. Atmos. Chem. Phys. 21 (22), 17079–17097.

Haidar, A., Verma, B.K., Haidar, R., 2019. A swarm based optimization of the XGboost parameters. Aust. J. Intell. Inf. Process. Syst. 16 (4), 74–81.

Holmes, R.M., Aminot, A., Kérouel, R., Hooker, B.A., Peterson, B.J., 1999. A simple and precise method for measuring ammonium in marine and freshwater ecosystems. Can. J. Fish. Aquat. Sci. 56 (10), 1801–1808.

Huang, G., 2021, February. Missing data filling method based on linear interpolation and lightgbm. In Journal of Physics: Conference Series (Vol. 1754, No. 1, p. 012187). IOP Publishing.

Huang, J., Kankanamge, N.R., Chow, C., Welsh, D.T., Li, T., Teasdale, P.R., 2018. Removing ammonium from water and wastewater using cost-effective adsorbents: a review. J. Environ. Sci. 63, 174–197.

Huang, H., Zhang, J., 2024. Prediction of chlorophyll a and risk assessment of water blooms in Poyang Lake based on a machine learning method. Environ. Pollut., 123501

Hussein, E.A., Thron, C., Ghaziasgar, M., Bagula, A., Vaccari, M., 2020. Groundwater prediction using machine-learning tools. Algorithms 13 (11), 300.

Icke, O., Van Es, D.M., de Koning, M.F., Wuister, J.J.G., Ng, J., Phua, K.M., Koh, Y.K.K., Chan, W.J., Tao, G., 2020. Performance improvement of wastewater treatment processes by application of machine learning. Water Sci. Technol. 82 (12), 2671–2680.

Ji, S.H., Baek, U.J., Shin, M.G., Goo, Y.H., Park, J.S. and Kim, M.S., 2019, September. Best feature selection using correlation analysis for prediction of bitcoin transaction count. In 2019 20th Asia-Pacific Network Operations and Management Symposium (APNOMS) (pp. 1-6). IEEE.

Jia, Z., Wang, J., Liu, X., Yan, Z., Bai, X., Zhou, X., He, X., Hou, J., 2023. Sediment diffusion is feasible to simultaneously reduce nitrate discharge from recirculating aquaculture system and ammonium release from sediments in receiving intensive aquaculture pond. Sci. Total Environ. 858, 160017.

Jones, R.D., Hood, M.A., 1980. Effects of temperature, pH, salinity, and inorganic nitrogen on the rate of ammonium oxidation by nitrifiers isolated from wetland environments. Microb. Ecol. 6, 339–347.

Kang, M. and Tian, J., 2018. Machine Learning: Data Pre-processing. Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things, pp.111-130.

Kanjilal, B., Masoumi, A., Sharifi, N., Noshadi, I., 2024. Ammonia harms and diseases: ammonia corrosion hazards on human body systems (liver, muscles, kidney, brain). In Progresses in Ammonia: Science, Technology and Membranes. Elsevier, pp. 307–324.

Khozani, Z.S., Banadkooki, F.B., Ehteram, M., Ahmed, A.N., El-Shafie, A., 2022. Combining autoregressive integrated moving average with Long Short-Term Memory neural network and optimisation algorithms for predicting ground water level. J. Clean. Prod. 348, 131224.

Khullar, S., Singh, N., 2021. Machine learning techniques in river water quality modelling: a research travelogue. Water Supply 21 (1), 1–13.

Kombo, O.H., Kumaran, S., Sheikh, Y.H., Bovim, A., Jayavel, K., 2020. Long-term groundwater level prediction model based on hybrid KNN-RF technique. Hydrology 7 (3), 59.

Krapac, I.G., Dey, W.S., Roy, W.R., Smyth, C.A., Storment, E., Sargent, S.L., Steele, J.D., 2002. Impacts of swine manure pits on groundwater quality. Environ. Pollut. 120 (2), 475–492.

Kube, M., Jefferson, B., Fan, L., Roddick, F., 2018. The impact of wastewater characteristics, algal species selection and immobilisation on simultaneous nitrogen and phosphorus removal. Algal Res. 31, 478–488.

Kushwaha, N.L., Kudnar, N.S., Vishwakarma, D.K., Subeesh, A., Jatav, M.S., Gaddikeri, V., Abdelaty, I., 2024. Stacked hybridization to enhance the performance of artificial neural networks (ANN) for prediction of water quality index in the Bagh river basin, India. Heliyon.

Li, S., Jin, X., Xuan, Y., Zhou, X., Chen, W., Wang, Y.X., Yan, X., 2019. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. Adv. Neural Inf. Process. Syst. 32.

Li, Y., Li, R., 2023. Predicting ammonia nitrogen in surface water by a new attention-based deep learning hybrid model. Environ. Res. 216, 114723.

Li, H., Liu, P., Guo, S., Ming, B., Cheng, L., Zhou, Y., 2018. Hybrid two-stage stochastic methods using scenario-based forecasts for reservoir refill operations. J. Water Resour. Plan. Manag. 144 (12), 04018080.

Li, T., Lu, J., Wu, J., Zhang, Z., Chen, L., 2022. Predicting aquaculture water quality using machine learning approaches. Water 14 (18), 2836.

Li, H.M., Tang, H.J., Shi, X.Y., Zhang, C.S., Wang, X.L., 2014. Increased nutrient loads from the Changjiang (Yangtze) River have led to increased harmful algal blooms. Harmful Algae 39, 92–101.

Liakos, K.G., Busato, P., Moshou, D., Pearson, S., Bochtis, D., 2018. Machine learning in agriculture: A review. Sensors 18 (8), 2674.

Liang, Y., Ma, R., Nghiem, A., Xu, J., Tang, L., Wei, W., Prommer, H., Gan, Y., 2022. Sources of ammonium enriched in groundwater in the central Yangtze River Basin: anthropogenic or geogenic? Environ. Pollut. 306, 119463.

Lim, B., Arık, S.Ö., Loeff, N., Pfister, T., 2021. Temporal fusion transformers for interpretable multi-horizon time series forecasting. Int. J. Forecast. 37 (4), 1748–1764.

Lin, J., Krom, M.D., Wang, F., Cheng, P., Yu, Q., Chen, N., 2022. Simultaneous observations revealed the non-steady state effects of a tropical storm on the export of particles and inorganic nitrogen through a river-estuary continuum. J. Hydrol. 606, 127438.

Lin, X., Li, X., Gao, D., Liu, M., Cheng, L., 2017. Ammonium production and removal in the sediments of Shanghai river networks: spatiotemporal variations, controlling factors, and environmental implications. J. Geophys. Res.: Biogeosci. 122 (10), 2461–2478.

Liu, X., Wang, J., 2023. Selective oxidation of ammonium to nitrogen gas by advanced oxidation processes: reactive species and oxidation mechanisms. J. Environ. Chem. Eng., 110263

Liu, X., Zhao, D., Xiong, R., Ma, S., Gao, W., Sun, H., 2011. Image interpolation via regularized local linear regression. IEEE Trans. Image Process. 20 (12), 3455–3469.

Ma, J., Li, P., Lin, K., Chen, Z., Chen, N., Liao, K., Yuan, D., 2018. Optimization of a salinity-interference-free indophenol method for the determination of ammonium in natural waters using o-phenylphenol. Talanta 179, 608–614.

Maganathan, T., Senthilkumar, S., Balakrishnan, V., 2020. Machine learning and data analytics for environmental science: A Review, prospects and challenges (November). In: In IOP conference series: materials science and engineering, 955. IOP Publishing, 012107 (November).

Maranon, E., Ulmanu, M., Fernandez, Y., Anger, I., Castrillón, L., 2006. Removal of ammonium from aqueous solutions with volcanic tuff. J. Hazard. Mater. 137 (3), 1402–1409.

Marcellino, M., Stock, J.H., Watson, M.W., 2006. A comparison of direct and iterated multistep AR methods for forecasting macroeconomic time series. J. Econ. 135 (1-2), 499–526.

Mejía, L., Barrios, M., 2023. Identifying watershed predictors of surface water quality through iterative input selection. Int. J. Environ. Sci. Technol. 20 (7), 7201–7216.

Murray, E., Treweek, S., Pope, C., MacFarlane, A., Ballini, L., Dowrick, C., Finch, T., Kennedy, A., Mair, F., O'Donnell, C., Ong, B.N., 2010. Normalisation process theory: a framework for developing, evaluating and implementing complex interventions. BMC Med. 8, 1–11.

Nalluri, M., Pentela, M., Eluri, N.R., 2020. A scalable tree boosting system: XG boost. Int. J. Res. Stud. Sci. Eng. Technol. 7 (12), 36–51.

Newhart, K.B., Marks, C.A., Rauch-Williams, T., Cath, T.Y., Hering, A.S., 2020. Hybrid statistical-machine learning ammonia forecasting in continuous activated sludge treatment for improved process control. J. Water Process Eng. 37, 101389.

Nollet, L.M. and De Gelder, L.S. eds., 2000. Handbook of water analysis. CRC press.

Ortiz-Santaliestra, M.E., Marco, A., 2015. Influence of dissolved oxygen conditions on toxicity of ammonium nitrate to larval natterjack toads. Arch. Environ. Contam. Toxicol. 69, 95–103.

Pajares, S., Ramos, R., 2019. Processes and microorganisms involved in the marine nitrogen cycle: knowledge and gaps. Front. Mar. Sci. 6, 739.

Parvathy, A.J., Das, B.C., Jifiriya, M.J., Varghese, T., Pillai, D., Rejish Kumar, V.J., 2023. Ammonia induced toxico-physiological responses in fish and management interventions. Rev. Aquac. 15 (2), 452–479.

Pei, S.T., Jiang, S., Liu, Y.R., Huang, T., Xu, K.M., Wen, H., Zhu, Y.P., Huang, W., 2015. Properties of ammonium ion–water clusters: analyses of structure evolution, noncovalent interactions, and temperature and humidity effects. J. Phys. Chem. A 119 (12), 3035–3047.

Perović, M., Šenk, I., Tarjan, L., Obradović, V., Dimkić, M., 2021. Machine learning models for predicting the ammonium concentration in alluvial groundwaters. Environ. Model. Assess. 26 (2), 187–203.

Pinson, P., Kariniotakis, G., 2004. On-line assessment of prediction risk for wind power production forecasts. Wind Energy.: Int. J. Prog. Appl. Wind Power Convers. Technol. 7 (2), 119–132.

Popovic, D., Sifrim, A., Davis, J., Moreau, Y., De Moor, B., 2015. Problems with the nested granularity of feature domains in bioinformatics: the eXtasy case. BMC Bioinforma. 16, 1–11.

Putatunda, S. and Rama, K., 2019, December. A modified bayesian optimization based hyper-parameter tuning approach for extreme gradient boosting. In 2019 Fifteenth International Conference on Information Processing (ICINPRO) (pp. 1-6). IEEE.

Qiao, Z., Sun, R., Wu, Y., Hu, S., Liu, X., Chan, J., 2020. Microbial heterotrophic nitrification-aerobic denitrification dominates simultaneous removal of aniline and ammonium in aquatic ecosystems. Water, Air, Soil Pollut. 231, 1–14.

Ransom, K.M., Nolan, B.T., Traum, J.A., Faunt, C.C., Bell, A.M., Gronberg, J.A.M., Wheeler, D.C., Rosecrans, C.Z., Jurgens, B., Schwarz, G.E., Belitz, K., 2017. A hybrid machine learning model to predict and visualise nitrate concentration throughout the Central Valley aquifer, California, USA. Sci. Total Environ. 601, 1160–1172.

Roelofs, R., Shankar, V., Recht, B., Fridovich-Keil, S., Hardt, M., Miller, J., Schmidt, L., 2019. A meta-analysis of overfitting in machine learning. Adv. Neural Inf. Process. Syst. 32.

Sawyer, J., 2008. Surface waters: Ammonium is not ammonia. In: Integrated Crop Management News, 4. Iowa State University, p. 21.

Shaikh-Mohammad, B.N. and Siddiqui, K., 2021, May. Random Forest Regressor Machine Learning Model developed for mental health prediction based on MHI-5, PHQ-9 and BDI scale. In Proceedings of the 4th International Conference on Advances in Science & Technology (ICAST2021).

Sun, L., 2020. Application and improvement of xgboost algorithm based on multiple parameter optimization strategy. In 2020 (December). 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE). IEEE, pp. 1822–1825 (December).

Tang, T., Jiao, D., Chen, T., Gui, G., 2022. Medium-and long-term precipitation forecasting method based on data augmentation and machine learning algorithms. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 15, 1000–1011.

Tyralis, H., Papacharalampous, G., Langousis, A., 2019. A brief review of random forests for water scientists and practitioners and their recent history in water resources. Water 11 (5), 910.

Vafaei, N., Ribeiro, R.A., Camarinha-Matos, L.M., 2018. Data normalisation techniques in decision making: case study with TOPSIS method. Int. J. Inf. Decis. Sci. 10 (1), 19–38.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems, 30.

Vega, M., Pardo, R., Barrado, E., Debán, L., 1998. Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. Water Res. 32 (12), 3581–3592.

Wang, Y., Ni, X.S., 2019. A XGBoost risk model via feature selection and Bayesian hyper-parameter optimization. arXiv Prepr. arXiv 1901, 08433.

Wang, X., Qiao, M., Li, Y., Tavares, A., Qiao, Q., Liang, Y., 2023. Deep-learning-based water quality monitoring and early warning methods: a case study of ammonia nitrogen prediction in rivers. Electronics 12 (22), 4645.

Wang, Z., Wang, Y., Jia, F., Zhang, F., Klimenko, N., Wang, L., He, Z., Huang, Z., Liu, Y., 2024. Spatiotemporal fusion transformer for large-scale traffic forecasting. Inf. Fusion 107, 102293.

Wang, L., Zhao, C., Liu, X., Chen, X., Li, C., Wang, T., Wu, J., Zhang, Y., 2021. Non-linear effects of the built environment and social environment on bus use among older adults in china: an application of the xgboost model. Int. J. Environ. Res. Public Health 18 (18), 9592.

Watson, P.A., Berner, J., Corti, S., Davini, P., von Hardenberg, J., Sanchez, C., Weisheimer, A., Palmer, T.N., 2017. The impact of stochastic physics on tropical rainfall variability in global climate models on daily to weekly time scales. J. Geophys. Res.: Atmospheres 122 (11), 5738–5762.

Wu, X., Dobriban, E., Ren, T., Wu, S., Li, Z., Gunasekar, S., Ward, R., Liu, Q., 2020b. Implicit regularization and convergence for weight normalization. Adv. Neural Inf. Process. Syst. 33, 2835–2847.

Wu, Z., Zhou, Y., Wang, H., 2020a. Real-time prediction of the water accumulation process of urban stormy accumulation points based on deep learning. IEEE Access 8, 151938–151951.

Wulfmeyer, V., Behrendt, A., Kottmeier, C., Corsmeier, U., Barthlott, C., Craig, G.C., Hagen, M., Althausen, D., Aoshima, F., Arpagaus, M., Bauer, H.S., 2011. The Convective and Orographically-induced Precipitation Study (COPS): the scientific strategy, the field phase, and research highlights. Q. J. R. Meteorol. Soc. 137 (S1), 3–30.

Yang, N., Zhang, C., Wang, L., Li, Y., Zhang, W., Niu, L., Zhang, H., Wang, L., 2021. Nitrogen cycling processes and the role of multi-trophic microbiota in dam-induced river-reservoir systems. Water Res. 206, 117730.

Yang, N., Zhang, C., Wang, L., Li, Y., Zhang, W., Niu, L., Zhang, H., Wang, L., 2021. Nitrogen cycling processes and the role of multi-trophic microbiota in dam-induced river-reservoir systems. Water Res. 206, 117730.

Zeng, F., Pan, Y., Yuan, X., Wang, M., Guo, Y., 2024. Transformer-based user charging duration prediction using privacy protection and data aggregation. Electronics 13 (11), 2022.

Zhang, G., Ruan, J., Du, T., 2020. Recent advances on photocatalytic and electrochemical oxidation for ammonia treatment from water/wastewater. Acs EsT. Eng. 1 (3), 310–325.

Zivot, E., Wang, J., Zivot, E., Wang, J., 2003. Rolling analysis of time series. Model. Financ. Time Ser. S-® 299–346.

Zounemat-Kermani, M., Batelaan, O., Fadaee, M., Hinkelmann, R., 2021. Ensemble machine learning paradigms in hydrology: a review. J. Hydrol. 598, 126266.