International Journal of River Basin Management

ISSN: 1571-5124 (Print) 1814-2060 (Online) Journal homepage: www.tandfonline.com/journals/trbm20

Advancing river water quality prediction: a comparative assessment of deep learning models for dissolved oxygen forecasting

Ali J. Ali & Ashraf A. Ahmed

To cite this article: Ali J. Ali & Ashraf A. Ahmed (16 Feb 2026): Advancing river water quality prediction: a comparative assessment of deep learning models for dissolved oxygen forecasting, International Journal of River Basin Management, DOI: 10.1080/15715124.2026.2626322

To link to this article: https://doi.org/10.1080/15715124.2026.2626322

Published online: 16 Feb 2026.

INTERNATIONAL JOURNAL OF RIVER BASIN MANAGEMENT
https://doi.org/10.1080/15715124.2026.2626322

Taylor & Francis
Taylor & Francis Group

Advancing river water quality prediction: a comparative assessment of deep learning models for dissolved oxygen forecasting

Ali J. Ali and Ashraf A. Ahmed

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

ABSTRACT

Accurate forecasting of dissolved oxygen (DO) is crucial for monitoring river water quality and protecting aquatic ecosystems. This study compares the performance of four deep learning models – Temporal Fusion Transformer (TFT), Informer, Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) – for forecasting DO concentrations in the River Lee (London, UK) across 7- and 30-day time frames. A multivariate time-series dataset was employed, with temperature, turbidity, pH, conductivity, chlorophyll, and river flow as predictors. Model skills were evaluated using RMSE, MAE, R2, and SMAPE. Over the 7-day period, TFT had the lowest RMSE (0.06) and SMAPE (8.86%), while LSTM had the greatest R2 (0.77). TFT outperformed Informer, LSTM, and GRU at the 30-day horizon, with R2 = 0.79 and SMAPE of 8.23%, despite significant accuracy losses. According to the variable contribution study, temperature and river flow were the most significant factors, particularly for short-term projections. Overall, the results show that transformer-based structures, particularly TFT, can successfully represent nonlinear temporal dependencies and multivariate interactions, making them ideal for multi-horizon DO forecasting in river systems. These models have the ability to supplement normal monitoring by offering short-term predictions about probable oxygen conditions.

ARTICLE HISTORY
Received 18 July 2025
Accepted 30 January 2026

ASSOCIATE EDITOR
Soufiane Haddout

KEYWORDS
Hydrological time series; River streamflow; Hydrological time series; River Lee; UK; Environmental monitoring; Multi-horizon prediction; Machine learning in water resources

1. Introduction

Dissolved oxygen (DO) is an essential parameter for assessing the condition and the quality of aquatic ecosystems. It is vital to the survival of aquatic life because it affects metabolic rates, the cycling of nutrients, and the overall equilibrium of ecosystems (Rajesh and Rehana 2022, Xiao *et al.* 2023). DO concentrations are significantly impacted by hydrological processes such as rainfall, river flow, and temperature variations, which are frequently caused by seasonal or climatic variability. Whilst slowing low-flow periods can cause oxygen depletion, especially in nutrient-rich water that supports algal blooms, high flows encourage aeration and oxygen diffusion (Luo *et al.* 2024). For the purpose of preventing hypoxic occurrences and protecting aquatic biodiversity, accurate DO prediction across short and medium time horizons is crucial.

While estimates between 7 and 30-days reflect medium-term perspectives pertinent to practical water-quality planning, forecasts within 7 days are often considered short-term (Ali et al., 2024; Ali and Ahmed 2024, 2025). Timely interventions like controlled flow releases or aeration control can be supported by accurate predictions across various scales (Danladi Bello *et al.* 2017, Deshpande *et al.* 2021). There have been recent developments in machine learning (ML) (Cox 2003, Kushwaha *et al.* 2024, Li *et al.* 2024, Pant *et al.* 2024). Model realism and contextual relevance are improved when hydrological elements like temperature, turbidity, and river flow are included in ML frameworks. Nevertheless, a lot of current methods have trouble capturing the nonlinear connections causing DO oscillations across hydrological regimes.

This study compares the performance of four sequential deep learning models – Temporal Fusion Transformer (TFT), Informer, Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) – in forecasting DO at 7- and 30-day timeframes. These models were chosen for their ability to represent temporal dependencies and variable interactions in multivariate hydrological data. This study attempts to determine if transformer-based designs outperform typical recurrent networks under dynamic river circumstances by focusing on the interface of hydrology and predictive modelling.

Hydrological and physicochemical elements combine to strongly control DO levels in river systems. One of the main drivers is river flow, which increases oxygenation through air exchange and turbulence

CONTACT Ali J. Ali ali.ali@brunel.ac.uk; Ashraf A. Ahmed ashraf.ahmed@brunel.ac.uk Department of Civil and Environmental Engineering, Brunel University London, Uxbridge UB8 3PH, UK

(Lamping *et al.* 2005, Haider *et al.* 2013). On the other hand, extended low conditions encourage oxygen depletion and stagnation, especially in nutrient-enriched reaches. Seasonal and climatic fluctuations are important regulators of DO dynamics because temperature has an inverse influence on DO solubility – warmer water stores less oxygen (Kulkarni 2016).

Additional elements that affect oxygen transport and ecosystem metabolism include turbidity, salinity, and conductivity (Irvine *et al.* 2011, El-Nahhal *et al.* 2021). Elevated turbidity and conductivity modify light penetration and ion exchange processes, whereas high salinity decreases oxygen solubility. Runoff, anthropogenic inputs, and rainfall events frequently cause these factors to co-vary, resulting in nonlinear feedback in DO behaviour. Sensor drift, missing data, or regional heterogeneity may also have an impact on their readings (Kim *et al.* 2021, Ghobadi *et al.* 2024).

Building reliable forecasting models requires an understanding of these connected interactions. Deep learning architectures, including TFT, Informer, LSTM, and GRU, may learn latent temporal interactions that conventional empirical formulations are unable to represent by incorporating such hydrological dependencies. In addition to supporting the more general objective of preserving riverine biological stability, this procedure increases forecasting accuracy.

Historical, physically based hydrological models that use deterministic equations to explain flow dynamics, temperature variation, and oxygen exchange have been used to estimate DO (Radwan *et al.* 2003, Pena *et al.* 2010). These models are often physically based, which means they rely on equations to represent physical processes in a body of water, such as flow dynamics, variation in temperature, and oxygen exchange mechanisms. Additionally, their efficiency over long or highly variable periods is limited since they require constant and high-resolution calibration data, which are not always accessible.

Machine learning and data-driven approaches have become more popular in recent years for water quality. These methods capture intricate temporal and spatial patterns by directly learning statistical relationships from observations without the need for explicit physical formulations (Li *et al.* 2024, Pant *et al.* 2024). A successful balance between interpretability and prediction ability has also been found in hybrid models that include data-driven and physical components (Xu *et al.* 2021, Ghobadi *et al.* 2024, Li *et al.* 2025). However, performance differs significantly between predicting horizons, model types, and data quality.

To overcome these constraints, this work develops and evaluates four deep learning architectures for a river system: Temporal Fusion Transformer (TFT), Informer, Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU). These models were chosen because of their shown capacity to incorporate hydrological variables as pH, turbidity, temperature, conductivity and river flow while modelling sequential and multi-horizon dependencies. The TFT uses gating mechanisms and multi-head attention to dynamically emphasise important variables (Lim *et al.* 2021, Maldonado-Cruz and Pyrcz 2024). By using its ProbSparse attention mechanism, the Informer increases computing efficiency for medium-term prediction (Zhou *et al.* 2021). While the GRU offers a more straightforward, computationally light recurrent baseline, the LSTM efficiently captures both short- and long-term dependencies in water-quality time series (Khozani *et al.* 2022).

Few studies have thoroughly assessed transformers' comparative performance versus recurrent models for DO prediction at several horizons, despite their increasing usage in hydrological forecasting. By examining how attention-based and recurrent architectures react to short- and medium-term hydrological variability, this work closes that gap and provides insight into their advantages and disadvantages for practical water-quality forecasting (Dey and Salem 2017). It is important to note that this work is unique in that it compares and interprets known deep learning architectures for short- to medium-term dissolved oxygen forecasting, rather than developing new model structures.

Previous research has mostly focused on immediate or event-driven forecasting, leaving a major vacuum in the credible short- to medium-term projections that environmental agencies require for effective water management. This work closes this gap by thoroughly comparing cutting-edge sequential models such as Temporal Fusion Transformer (TFT), Informer, Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) over 7- and 30-day forecasting horizons. Furthermore, the incorporation of hydrological factors such as temperature, turbidity, pH, and river flow guarantees that the models are adapted to the intricate dynamics of aquatic environments. The key innovations of this study are:

- Examining the influence of environmental variation by illustrating how temperature, turbidity, and river flow affect model predictions by examining their effects on DO levels.
- Evaluating the capacity of sophisticated deep learning models (TFT, Informer, LSTM, and GRU) to properly anticipate DO levels in river systems across short (7-day) and medium (30-day) time periods.
- In-depth analysis of hydrological factors – such as pH, chlorophyll, turbidity, temperature, conductivity, and river flow – affects these models' predictions and how they relate to managing the water quality of rivers.

- Cross-site validation involves testing the models at a different monitoring site to assess the created frameworks' spatial transferability and resilience.

This study contributes to our knowledge of how deep learning models incorporate hydrological variability throughout operational forecasting timeframes by comparing attention-based and recurrent architectures. The results provide quantifiable information on how well-suited contemporary transformer frameworks are for short-term dissolved oxygen forecasting in river systems.

2. Materials and methods

This section highlights our approach for Dissolved oxygen (DO) forecasting in the River Lee. It uses a combination of transformer models and deep learning models, including TFT, Informer, LSTM, and GRU. The procedures for gathering and preparing data are presented in this section, outlining the mechanics of each machine learning model, and clarifying the evaluation standards.

2.1. Data collection and pre-processing

In order to forecast dissolved oxygen (DO) levels, this study uses a wealth of hydrological data from the River Lee in East London, United Kingdom. The dataset, which comes from the Environment Agency's Hydrology Data Explorer, covers the period from March 2016 to January 2024. The monitoring station is about 7 kilometres from the River Thames, towards East London. The dataset contains daily observations generated from hourly readings from March 2016 to January 2024. To guarantee comparability across inputs, all variables were normalised before model training. Dissolved oxygen, an important indication of river health, was analysed with other major water quality indicators such as pH, chlorophyll, turbidity, temperature, conductivity, and river flow. Data were acquired from the same monitoring station in the river. However, the availability of comprehensive data for all characteristics differed across the stations that were located in the area. As a result, The River Lee was chosen as a model urban river system due to human stresses, changing hydrological conditions, and known water-quality issues. Rather than focussing on unique hydrodynamic behaviour, the work use the River Lee as a realistic testbed for evaluating the performance and interpretability of deep learning models under typical urban river settings where short-term dissolved oxygen forecasting is operationally important.

An additional monitoring station close to the River Lee was utilised as an external validation site in order to evaluate spatial resilience. The additional dataset allows for initial testing of model transferability under various hydrological conditions, although covering a shorter time period and fewer variables (Appendix A).

Raw hourly readings were aggregated to daily averages to smooth out sub-day noise and highlight hydrological interactions that influence DO. Missing values (<2%) were filled via mean imputation, ensuring sequence continuity for deep-learning models (Vafaei *et al.* 2018). Sensitivity testing revealed that the results were consistent before and after imputation. Outliers, primarily from sensor spikes, were identified using interquartile range criteria and ocular inspection. All predictors were normalised to [0, 1] by the MinMaxScaler (Kang and Tian 2018). Previous research has linked temperature, river flow, turbidity, conductivity, pH, and chlorophyll to oxygen solubility and water clarity (Neal *et al.* 2006, Kney and Brandes 2007, Huey and Meyer 2010). To guarantee repeatability, data preparation and modelling were carried out in Python using TensorFlow and scikit-learn, as well as a fixed random seed. To ensure stability, the findings were cross-checked with different random splits and minor changes to imputed data. A schematic flowchart of data collecting, preprocessing, model training, validation, and assessment.

Several preprocessing and feature-engineering methods were used to prepare the dataset for dissolved oxygen (DO) predictions. Key water-quality indicators – DO, pH, chlorophyll, turbidity, temperature, conductivity, and river flow – were kept recording the physical and chemical variables that influence oxygen solubility and ecosystem metabolism. Temperature directly influences oxygen solubility (Zhi *et al.* 2023), whereas turbidity shows suspended particles that decrease light penetration and photosynthetic activity.

Lagged features were created for all predictors except DO with a three-day lag, allowing models to learn temporal relationships between past and current situations (Kim *et al.* 2021). Rolling averages with a seven-day timeframe were calculated to smooth daily swings and capture weekly hydrological patterns (Amor *et al.* 2016). Together, these modifications provide the contextual memory required for sequential deep-learning systems. To equalise the contribution of each input feature, all variables were scaled from 0 to 1 using the MinMaxScaler (Deepa and Ramesh 2022, Ahmed *et al.* 2024). Scaling keeps large-magnitude factors like conductivity from dominating gradient updates in models like as LSTM, GRU, and Transformer versions.

A multi-horizon dataset was subsequently created for 7- and 30-day prediction windows, allowing short- and medium-term forecasting. The dataset was separated into three subsets: training (70%), validation (15%), and holdout testing (15%). Sampling was random yet reproducible, using a set random

seed (42). Although chronological divides are usual in time-series research, random sampling was used since the dataset had minimal autocorrelation beyond 30 days and rather steady hydrological behaviour, resulting in balanced distributions across subsets (Lessels and Bishop 2020).

The training and validation sets were used to fit models and tune hyperparameters, with the unseen test set serving as the final evaluation. The uniform preprocessing and feature-engineering approach guaranteed that all models – TFT, Informer, LSTM, and GRU – were trained with similar inputs, allowing for a fair multi-model comparison. The multi-horizon approach allowed for simultaneous calculation of near-term (7-day) and medium-term (30-day) DO dynamics, which improved the interpretability of short- and long-term responses to hydrological variability (Quaedvlieg 2021).

2.2. Machine learning models

Input factors were chosen based on their known physical and biogeochemical effects on dissolved oxygen dynamics. Temperature determines oxygen solubility, turbidity and chlorophyll indicate light availability and biological activity, river flow affects mixing and reaeration, conductivity depicts ionic composition, and pH reflects chemical conditions that impact oxygen processes.

This study employs a novel method in the field of hydrological modelling by applying the Informer and TFT models to the challenge of DO forecasting across the specified (short 7-days) and medium (30-days). These structures are ideal for hydrological time series forecasting because they can capture nonlinear dependencies and long-term temporal interactions between many environmental causes. Unlike traditional hybrid or ensemble methods, which integrate many model types to improve accuracy, transformer-based systems learn temporal and contextual correlations directly from data using their attention processes (Choi and Lee 2023). To achieve total repeatability, all model designs, hyperparameters, preprocessing methods, and assessment procedures are thoroughly described. Fixed random seeds, consistent data splits, and identical feature sets were used to train the models across all architectures. This defined methodology assures that performance discrepancies are due to model capacity rather than experimental setting.

Transformers were chosen because their self-attention mechanism enables the simultaneous modelling of both short- and medium-range interdependence across all hydrological variables – features that recurrent networks frequently approximate sequentially. The informer uses a ProbSparse attention method to effectively handle large sequences while preserving accuracy (Zhou *et al.* 2021). In contrast, the TFT uses variable selection networks and gated residual connections to provide interpretable multi-horizon predictions (Lim *et al.* 2021)

For a comparative study, LSTM and GRU models were used as recurrent baselines, allowing for direct comparison of transformer performance to proven sequential architectures. While hybrid or physics-informed frameworks can improve interpretability, they usually need considerable preprocessing and parameter calibration (Sseguya and Jun 2024). In contrast, the transformer-based models utilised here dynamically prioritise important features from raw inputs, eliminating the need on handmade feature design.

Although this work focuses on data-driven modelling, the ability to integrate transformer topologies with physics-based techniques remains an essential area of future research. Such integration might combine the interpretability of mechanistic models with the predictive ability of deep learning, hence increasing the application of DO forecasting over a wide range of hydrological systems.

2.2.1. Long short-term memory

This study used LSTM networks to forecast dissolved oxygen levels in a river system. Because of its capacity to store information over lengthy time periods, LSTM models are ideal for time series forecasting. They do this with memory blocks made up of gates and cells, with the cell serving as a conveyor belt to transport information along the sequence. The gates define what information should be added or deleted from memory, and each gate has its own weights and biases (Khozani *et al.* 2022, Ali and Ahmed 2024).

$$f_t = \sigma(\omega_f.[h_{t-1}, x_t], b_f)$$

The forget gate decides which information to erase from memory, where f_t represents the forget gate output, σ is the sigmoid function, h_{t-1} is the hidden state from the previous time step, x_t is the current input, ω_f is the forget gate's weight matrix, and b_f is the bias term. This procedure guarantees that the model preferentially remembers significant information, with the sigmoid function determining the degree of memory retention (Kong *et al.* 2021).

The input gate, which selects what new information should be placed in the memory cell, employs the tanh function to produce fresh data for the memory (Gundu and Simon 2021). These gates enable the model to dynamically determine which components of the time series are relevant for predicting. In the current study, LSTM model architecture was defined as follows:

- Model definition: A sequential model was used with an LSTM layer containing 64 units. The layer identifies temporal connections in the data without

returning the sequence. The model concludes with a dense layer that forecasts the multi-horizon sequence.
- Output layer: to anticipate dissolved oxygen levels across a number of future time steps, a dense layer with a linear activate function was employed.
- Model compilation: The model is trained well by using the Adam optimiser with a learning rate of 0.001 and a Mean Squared Error (MSE) loss function during compilation.
- Training and validation: Early stopping was utilised during training to prevent both overfitting and underfitting, halting the process if there are no enhancements in the validation loss after 10 epochs.

The model's performance is assessed by making predictions on the holdout, validation, and training sets after training. The gating mechanism of LSTM enables it to efficiently manage partial and noisy water quality data by filtering out unnecessary information (Khozani *et al.* 2022). It is perfect for estimating DO because of its ability to simulate non-linear interactions between environmental parameters. Furthermore, multi-horizon forecasting is supported by LSTM, enabling prediction across a longer time horizon. Previous hydrological studies have effectively used this model (Dayal *et al.* 2024, Martín-Suazo *et al.* 2024), demonstrating its ability to forecast intricate river dynamics. Figure: Workflow of the methods used to anticipate dissolved oxygen (DO) levels in the River Lee. The process consists of data collection, preprocessing, model building, training and validation, and performance evaluation (Figure 1).

2.2.2. Gated recurrent unit

GRU is a great option for sequence modelling because of its effectiveness in managing long-term dependencies, which is essential for forecasting hydrological variables like DO in river ecosystems (Chung *et al.* 2014). By removing the memory cell and integrating the input and forget gates into a single update gate, the GRU model of recurrent neural networks simplifies its design more than the LSTM. This maintains the capacity to accurately describe sequential data while reducing computing complexity (Cahuantzi *et al.* 2023).

To ensure the network can capture the dependencies, the update gate $\mathcal{Z}_t$ in the GRU regulates how much data from the previous time steps is carried over to the current state. The amount of historical data that should be forgotten is decided by the reset gate r_t. This can be expressed mathematically as (Chung *et al.* 2014):

$$\mathcal{Z}_t = \sigma(W_z x_t + U_z h_{t-1})$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1})$$

In this study, a GRU model was created to anticipate DO levels over a variety of time periods. The capacity of the GRU can handle intricate temporal linkages in hydrological data, making it a well-suited model for this research. For a fair comparison, the GRU model was constructed similarly to the LSTM model with 64 units, and a dense layer with a linear activation function to provide predictions for the designated horizon.

2.2.3. Temporal fusion transformer

The TFT model is a cutting-edge deep learning model built for time series forecasting, particularly multi-horizon applications such as predicting DO in river systems (Marcellino *et al.* 2006, Lim et al., 2021). TFT mixes standard model interpretability with deep learning's advanced sequence modelling capabilities. Its main strength is its capacity to handle complicated, non-linear temporal patterns found in hydrological data by using methods like multi-head self-attention and gated residual network (GRN). The multi-head self-attention mechanism assists the model in capturing long-term dependencies in time series data, whilst the GRN filters out unnecessary information, allowing the model to focus on the most important features (Vaswani *et al.* 2017).

Furthermore, TFT incorporates both static and temporal variables, such as geological characteristics and water flow, making it useful for forecasting river quality over long periods of time. The model analyses data using both LSTM layers, which capture short- and medium-term dependencies, and the attention layer, and is critical for comprehending patterns such as seasonal changes or lengthy dry spells that affect dissolved oxygen in river systems. The model includes early stopping to prevent overfitting and ensure resilience over many horizons. It is trained using the Adam Optimiser for efficient learning.

The TFT model is ideal for this study since it can handle multi-horizon predictions whilst maintaining interpretability, and it has been applied in some studies (Wang and Tang 2023). The nature of river systems, in which numerous environmental conditions impact DO, necessitates a model that can incorporate both short-term fluctuation, such as daily river flow changes, and long-term seasonal trends like temperature fluctuation. The TFT design, which incorporates multi-horizon attention and gating techniques, ensures the model captures the essential hydrological dynamics required for reliable prediction over a range of time horizons. The gating mechanism blocks can be expressed as follows (Lim *et al.* 2021):

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

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

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

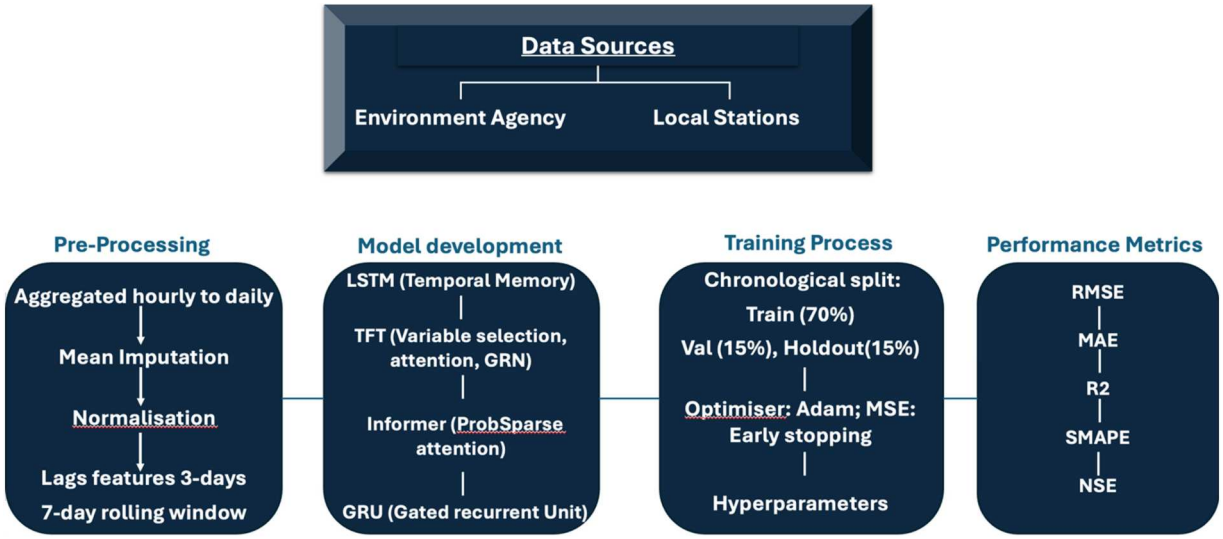

Figure 1. Workflow of the methods used to anticipate dissolved oxygen (DO) levels in the River Lee. The process consists of data collection, preprocessing, model building, training and validation, and performance evaluation.

With a primary input of a and an optional context vector of c, the Gated Residual Network (GRN) formula makes use of Gated Linear Units (GLUs) and exponential linear unit (ELU) activation (Clevert *et al.* 2015). A thorough grasp of the variables affecting river water quality is made possible by its potent temporal fusion decoder and capacity to analyse both static and dynamic covariates.

2.2.4. Informer

The Informer model was used in this research because it effectively manages long sequence time series forecasting, which is crucial for predicting DO levels in rivers. It solved significant constraints of standard transformer models by including innovations such as the ProbSparse self-attention mechanism and self-attention distilling operation, making it suited for such forecasts. The ProbSparse self-attention mechanism enhances computational efficiency by focusing on the most critical queries for self-attention computation, lowering complexity from $Q(L^2)$ to $Q(L \log L)$, where L is the length of the input sequence (Zhao and Wang 2023). The attention mechanism is determined by (Zhou *et al.* 2021):

$$Attention(Q,\ K,\ V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Where $Q,\ K,\ and\ V$ stand for the queries, keys, and values matrices, respectively, whilst d_k denotes the key's dimension. The ProbSparse approximation in the Informer model only chooses the top- u queries – where u is defined by the sparsity ratio – for calculating the self-attention. The self-attention distilling procedure decreases computing efforts by gradually lowering the input size over numerous levels. The relevance of each attention score is reduced, leaving just the most important components for succeeding (Zhou *et al.* 2021, Xu *et al.* 2023). It can be expressed as:

$$\hat{A}_{i,\,j} = \frac{A_{i,j}}{\sum_{k=1}^{L} A_{i,\,k}}$$

Where $\hat{A}_{i,\,j}$ is the condensed form, and A is the original attention matrix. While preserving important temporal relationships, this procedure efficiently condenses the sequence information to make it easier to handle for deep processing.

Finally, the model employs a generative decoder, which allows it to predict a whole future sequence in a single forward pass rather than step by step, hence boosting both inference speed and prediction accuracy for extended sequences. The model used in this study used 4 attention heads with a dimensionality of the model's embedding space of 64 and a feed-forward dimension of 128. The ProbSparse self-attention mechanism decreases the computational complexity, which leads to more efficiency in long-sequence prediction over extended horizons. The Informer model's capacity to accommodate extended input sequences while focusing on crucial temporal dependencies makes it ideal for forecasting dissolved oxygen in rivers, where seasonal and hydrological trends play an important role.

These models were chosen for their interpretability features, which are essential for evaluating our hypothesis, in addition to their forecasting accuracy. The attention processes included in the informer and TFT models are crucial for understanding how different hydrological conditions affect DO levels. For instance, it's clear which variables (such as temperature or river flow) the TFT model concentrates on under various circumstances or at different seasons of

the year due to its multi-head attention mechanism (Lim *et al.* 2021). This skill helps to design focused management plans by explicitly testing our hypothesis that some environmental elements have a more substantial influence during various periods. Water management authorities can better comprehend model projections and make well-informed judgments because of these interpretability qualities, which also make practical applications easier. Authorities can improve water quality management methods by better prioritising monitoring and intervention efforts by determining which factors have the most impact on DO levels.

2.2.5. Hyperparameter configuration

Each model's hyperparameters were chosen with care to maximise performance through a range of predicting horizons. A batch size of 32, a learning rate of 0,001, and 64 units in the hidden layers were chosen for the LSTM and GRU models. To avoid overfitting, both models used the Adam Optimiser and early stopping with a 10-epoch patience. Crucial hyperparameters for the TFT model were two 64- and 32-unit LSTM layers, followed by a multi-head attention mechanism with four heads and a key dimension of 16. Four attention heads, a sparse attention mechanism, an embedding size of 64, and a feed-forward dimension of 128 were all included in the informer model's configuration. The Adam Optimiser was used to train both the informer and TFT models with a batch size 32, an early stopping time of 10 epochs, and a learning rate of 0.001. Taking into account the distinct temporal and hydrological dynamics of the dataset, these hyperparameters were adjusted to strike a balance between accuracy and computing efficacy.

2.3. Evaluation methods

2.3.1. Rolling window

Each model in our study was assessed using a standardised procedure designed to forecast the amounts of dissolved oxygen in river ecosystems. During post-data pre-processing, a rolling window technique that produces rolling window and lag features was used. This method was used to capture the impact of historical environmental circumstances on present-day dissolved oxygen levels. By using information from previous observations (i-1 to i-N), the rolling window approach evaluates the model's performance at a specific time instance i to produce h-step forward predictions (Amor *et al.* 2016). This method is in line with the dynamic character of river habitats, where past circumstances influence the current water quality standards, such as variations in temperature, river flow, or turbidity. As the time window advances, the model is updated regularly to take into account fresh data and enhance future projections. Our estimates will always be precise and true to the natural fluctuations in river systems, thanks to this continuous updating mechanism.

2.3.2. Holdout sets technique

In this study, the holdout method was utilised to assess how well our models performed on forecasting DO levels. This method is particularly effective since it splits the dataset into two separate sets – a training set and a holdout (testing) set. This method works especially well for non-stationary time series, such as data on river water quality. The models' predictive power was assessed by training them on the first segment and testing them on the second, which included new, unseen data. Compared to other approaches, such as cross-validation, the holdout technique is the most appropriate for time-dependent data, as it assures more reliable validation, which makes it particularly beneficial for hydrological forecasting. It provides a reliable evaluation of the model's predictive power in a practical setting, making it an important stage in confirming its performance (Cerqueira *et al.* 2020).

2.3.3. Performance metrics (RMSE, MAE, R^2, and SMAPE)

Three metrics were used to assess the effectiveness of our models: the Root Mean Square Error (RMSE), Mean Absolute Error (MAE), R-squared (R^2), and the Symmetric Mean Absolute Percentage Error (SMAPE). These metrics were selected to provide a comprehensive evaluation of the prediction accuracy and reliability of the model (Chicco *et al.* 2021, Li *et al.* 2025).

The RMSE determines the average error magnitude by assessing the extent of anticipated errors. The accuracy of the model is determined by taking the square root of the average squared differences between the actual and projected values. The RMSE formula is as follows:

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

MAE determines the average magnitude of errors in a set of forecasts without taking into account the direction of the errors. It is the mean, with equal weight given to each individual deviation, of the absolute differences between the observed and predicted values over the test sample. The formula can be expressed as follows:

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

R-squared (R^2) measures the dependent variable's

anticipated variation from the independent factors. Based on the percentage of total variation that the model explains, it evaluates how well observed results are repeated by the model in DO levels. R^2 can be calculated as follows:

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

SMAPE is popular error static in time series models, the accuracy is determined by comparing the expected and actual values and normalising the absolute errors by the total of the absolute values of the anticipated and actual values. This is especially important in ecological research because DO concentrations can fluctuate greatly. SMAPE can be expressed as follow:

$$SMAPE = 100\ \times mean\left(\frac{2 \times |y_{prep} - y_{true}|}{|y_{true}| + |y_{prep}|}\right);$$
$$(best\ value = 0\%,\ \ worst\ value = 100)$$

To handle the unique difficulties presented by DO level forecasts in river systems, the forecasting horizons were carefully chosen in addition to the performance metrics. Numerous short- and long-term environmental variables impact the temporal dynamics of DO levels. In order to fully capture these dynamics, our analysis incorporates a variety of predicting horizons, including 7 and 30.

- Short-term horizons (7 and 30 days): These time periods are essential for comprehending how DO levels react instantly to fleeting occurrences like precipitation, pollution, and abrupt ecological shifts. The survival of aquatic species and the health of ecosystems may depend on prompt responses, which require management of these transient changes.

The choice of these particular horizons enables our models to offer insightful information at a variety of temporal scales, each of which is essential for distinct facets of ecological forecasting and water quality management. It is guaranteed that our prediction models are not only adaptable but also directly relevant to requirements of environmental scientists and local government agencies in maintaining the well-being of river systems by including this range.

Table 1. The holdout results of all models across all 5 horizons.

	7-Step Ahead (Holdout)			
Model	RMSE	MAE	R^2	SMAPE
TFT	0.06	0.05	0.77	**8.86**
Informer	0.07	0.05	0.74	**9.14**
LSTM	0.06	0.05	0.77	**9.07**
GRU	0.06	0.05	0.76	**9.10**
	30-Step Ahead (Holdout)			
Model	RMSE	MAE	R^2	SMAPE
TFT	0.06	0.04	0.79	8.23
Informer	0.07	0.05	0.73	9.28
LSTM	0.06	0.05	0.75	8.87
GRU	0.07	0.05	0.70	9.84

3. Results and discussion

In this section, four model was assessed – TFT, Informer, LSTM and GRU – over two predicting horizons: 7 and 30 days. This multi-horizon forecasting technique reveals important information about each model's resilience, accuracy and capacity to manage short- and medium-term dependencies in predicting DO levels in river water. The model's performance was evaluated using RMSE, MAE, R^2 and SMAPE. These metrics examine the models' capacity to forecast DO and capture trends and variability in water quality. A complete examination of how each model performs over the various horizons is presented, with an emphasis on the identification of standout models for each horizon. The fundamental contribution of this work is its comparative and interpretive analysis, rather than architectural alteration. The study evaluates transformer-based and recurrent deep learning models under identical hydrological inputs and forecasting horizons, providing practical insight into the relative strengths, stability, and interpretability of each architecture for short- to medium-term dissolved oxygen forecasting. Table 1 demonstrate the results on the holdout sets.

In the medium-term, TFT had the best performance (RMSE = 0.06; SMAPE = 8.23; R2). Over the near run, all models had comparable RMSE (0.06-0.07) and MAE (0.05). TFT had the lowest SMAPE (8.86%) and highest R2 (0.77), closely followed by LSTM and GRU (SMAPE 9.07-9.10). Informer showed somewhat larger inaccuracy (SMAPE = 9.14%). These findings show that all designs caught short-term DO changes successfully, with TFT having the most consistent error profile.

This study compares the performance of four sequential deep learning models – Temporal Fusion Transformer (TFT), Informer, Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU)

Table 2. The validation results of all models across all 5 horizons.

	7-Step Ahead (Validation)			
Model	RMSE	MAE	R^2	SMAPE
TFT	0.07	0.05	0.73	**9.35**
Informer	0.07	0.05	0.73	**9.64**
LSTM	0.07	0.05	0.74	**9.57**
GRU	0.07	0.05	0.74	**9.50**
	30-Step Ahead (Validation)			
Model	RMSE	MAE	R^2	SMAPE
TFT	0.06	0.05	0.77	9.33
Informer	0.07	0.05	0.70	10.35
LSTM	0.07	0.05	0.70	10.28
GRU	0.08	0.06	0.65	11.25

– in forecasting DO at 7- and 30-day timeframes. These models were chosen for their ability to represent temporal dependencies and variable interactions in multivariate hydrological data. This study attempts to determine if transformer-based designs outperform typical recurrent networks under dynamic river circumstances by focussing on the interface of hydrology and predictive modelling.

Strong generalisation across all models was confirmed by the validation results (Table 2), which closely matched the holdout findings. SMAPE over the seven-day horizon varied from 9.35% to 9.64%, with TFT once more yielding the lowest error (9.35%). Informer's findings were somewhat better than those of LSTM and GRU. TFT continued to perform the best for the 30-day horizon (SMAPE = 9.33%). GRU demonstrated lower accuracy (SMAPE = 11.25%), showing more difficulties modelling longer-term temporal patterns, whereas Informer and LSTM had slightly higher error (10.28–10.35%).

Early stopping, preprocessing, and feature-engineering methods successfully prevented overfitting, and the models reflected robust hydrological connections in the data, as seen by the strong relationship between validation and holdout measures.

3.1. Model overview

The Symmetric Mean Absolut Percentage Error (SMAPE) was used as a main accuracy metric to evaluate the performance of four prediction models used in this study: TFT, Informer, LSTM, and GRU, over a range of predicting horizons. The architecture of each model is specifically designed to capture temporal relationships, which affects how successful they are at various time scales. Figure 2 summaries the holdout SMAPE performance for both horizons.

3.1.1. 7 – steps forecast

The substantial temporal autocorrelation in short-term DO changes was reflected in all models' good performance over the 7-day horizon. While Informer, LSTM, and GRU generated similar SMAPE values (9.07–9.14%), TFT had the lowest error (8.86%). This shows that all models successfully captured daily DO fluctuation, with just a little difference in predicted accuracy.

3.1.2. 30 – steps forecast

Performance patterns remained consistent throughout 30 days, although discrepancies across models grew more prominent. TFT again had the lowest error rate (8.23%), followed by LSTM (8.87%). Informer and GRU had greater SMAPE (9.28% and 9.84%), indicating that their capacity to extract medium-range temporal relationships was less effective than TFT's variable-selection and attention processes.

Each model's actual vs expected graphs over a range of timeframes offer important information about the models' strengths and weaknesses. Figure 3 plots emphasise each model's accuracy and reactivity to environmental changes by graphically illustrating how it depicts the dynamics of dissolved oxygen in the river system over time.

For the 7 days horizon, both the TFT and informer models aligned well with the real trend in DO, indicating how well they manage transient variations. The models' susceptibility to sudden changes in water quality parameters, including rainfall or pollution influxes, could be the cause of the minor differences between expected and actual readings. By using their advanced attention processes to concentrate on the most important recent future, as indicated by the denser clusters of points surrounding the line perfect fit. The same goes for the LSTM and GRU, which operated admirably. However, the graph shows occasional departures during periods of significant DO oscillations, which may indicate a worse ability to respond to quick environmental changes than transformer-based models. The error distribution in these models is slightly wider, indicating a marginally worse prediction precision at this scale; this can be viewed in the next section.

3.2. Model performance

Error-distribution plots (Figures 3–6) were used to evaluate each model's stability and bias across the 7- and 30-day periods. These distributions offer more insight into predicting dependability than aggregate measurements like RMSE or SMAPE (Mertikas, 2023).

Figure 4 illustrates that the TFT generated error distributions that were tightly centred and very symmetrical across both horizons. At 7 days, errors were tightly grouped around zero, showing high short-term stability and low systematic bias. Although the dispersion grew somewhat at 30 days, the distribution remained centred with a tiny mean error (≈0.01), indicating that TFT performed consistently in the medium range. This conduct is consistent with the model's attention processes, which aid in the preservation of relevant temporal structure over long time horizons.

The Informer model (Figure 5) likewise provided well-centred error distributions on both horizons. The 7-day distribution was compact, demonstrating the model's capacity to capture local temporal relationships. At 30 days, the error spread rose considerably, while the median error remained close to 0. This expansion is predicted because longer time horizons create more uncertainty and amplify modest departures in earlier phases. Informer's stability across horizons reflects the advantages of its sparse attention mechanism, while its medium-range mistakes were somewhat bigger than TFT's.

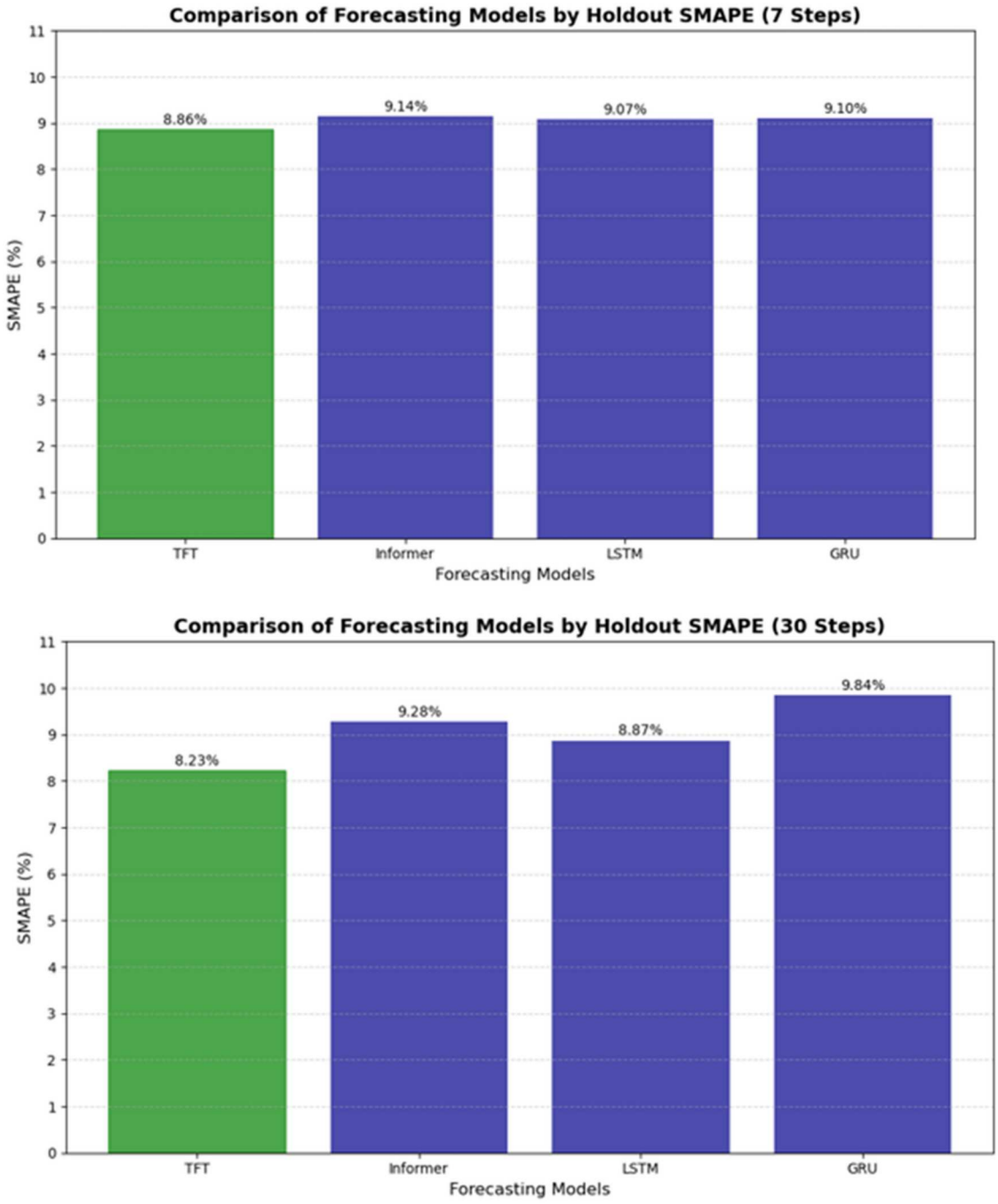

Figure 2. Comparison between the models in the short-horizons.

At 7 days, LSTM demonstrated consistent short-term behaviour, with a tight, central distribution (Figure 6). Over the 30-day period, the dispersion widened more than the transformer-based models. This shows that, while LSTM captures short-term hydrological connections successfully, it loses robustness as the forecast window grows larger. This is consistent with recurrent models' reliance on sequential memory, which can accrue errors over extended time periods.

GRU showed similar performance trends as LSTM (Figure 7). The 7-day horizon revealed a tight distribution at zero, showing high short-term dependability. At 30 days, the distribution expanded and exhibited somewhat greater deviations than LSTM, indicating that GRU's simpler gating structure had a limited capacity to sustain medium-term dependencies. Nonetheless, the median error remained close to zero, showing no systematic bias.

Across all models, error distributions were tightest at the 7-day horizon, which is consistent with the high short-term temporal autocorrelation commonly seen in DO time series. At the 30-day horizon, all models had larger distributions, indicating the predicted rise in uncertainty as hydrological factors pile over time. TFT and Informer had the narrowest error distributions across both horizons, showing higher medium-range stability, which is likely due to their attention processes and capacity to focus on informative time steps.

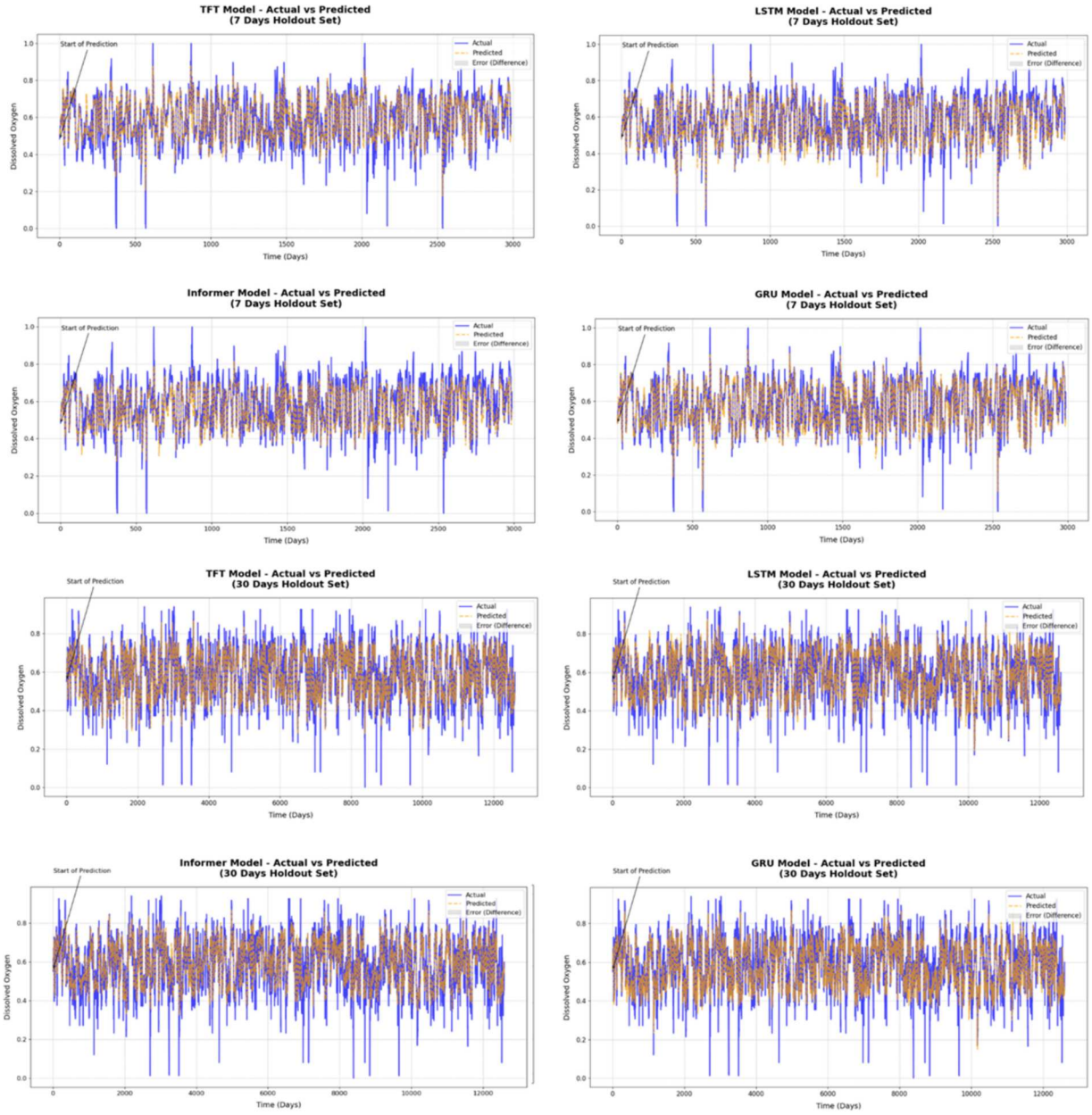

Figure 3. actual vs predicted for short term (7- and 30-steps).

In contrast, LSTM and GRU had a wider variety of mistakes at 30 days, indicating the difficulty recurrent architectures have in maintaining information over long durations. When combined with the aggregate accuracy measures, these patterns indicate that transformer-based models conserve temporal structure more successfully over long horizons, whereas recurrent models remain competitive but are less stable in medium-term forecasting.

3.3. Variable contribution analysis

The major goal variable in this work was dissolved oxygen (DO), and its linkages to important hydrological and environmental parameters give critical context for understanding model behaviour. As seen in Figure 8, temperature had the highest negative correlation with DO (−0.69). This matches well-known thermodynamic phenomena, in which oxygen solubility decreases as water temperature rises. Models that properly captured this link, notably TFT, produced lower short-horizon SMAPE values (8.86% at 7 days and 8.23% at 30 days), demonstrating the importance of incorporating quick temperature-driven oscillations for prediction accuracy. At these shorter horizons, the LSTM model also demonstrated impressive accuracy, successfully monitoring quick changes in water metrics as turbidity, pH, and chlorophyll. LSTM capitalised on the temporal correlation between DO and pH (correlation of 0.13), especially in reaction to short-term acidity variations, whilst having a simpler recurrent structure than attention-based models. This slight direct association, however, indicates that its impact decreased after the short-term projection periods.

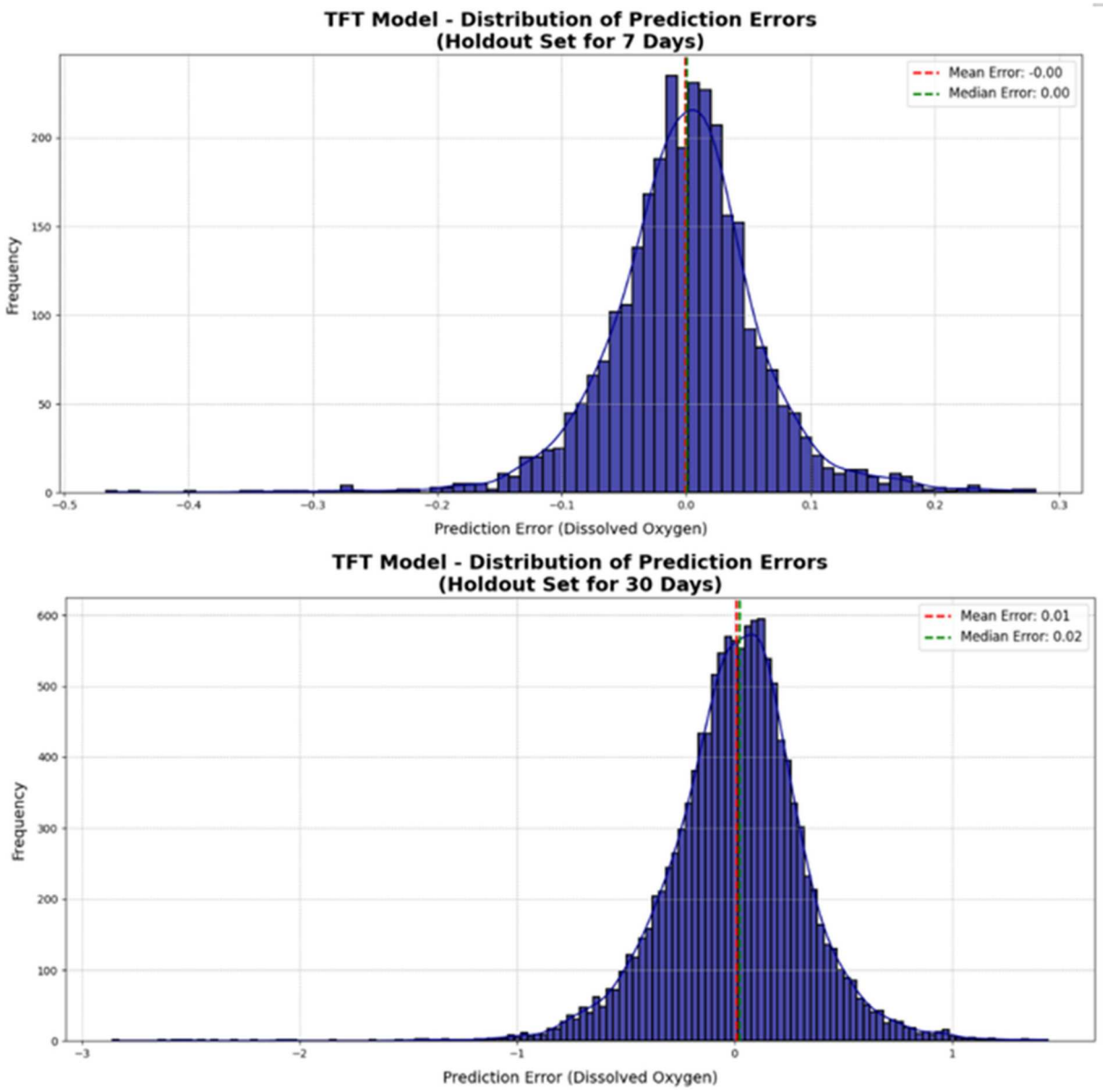

Figure 4. TFT error distribution for both horizons.

LSTM also worked well at short time scales, successfully predicting variations related to pH, turbidity, and chlorophyll. The positive but slight correlation between DO and pH (0.13) indicates a weaker direct influence, which may explain why the effect was more visible in short-term forecasts than over longer time horizons. Similarly, chlorophyll had a slight positive connection with DO (0.20), indicating the impact of biological activity. This link contributed to the successful capture of short-term changes by LSTM and TFT, while Informer and GRU were less sensitive to these fast swings.

GRU handled short-term dependencies consistently and provided performance equivalent to LSTM at both horizons. Turbidity showed a moderate association with river flow (0.53), indicating that hydrological disturbances influence suspended particles. Both GRU and LSTM appeared sensitive to this relationship over a 7-day period, but GRU's error distribution widened more noticeably at 30 days. This implies that cumulative nonlinear interactions among hydrological factors affected its medium-term stability.

Informer worked well on short time horizons while being meant for longer input sequences. Its multi-head attention mechanism allowed it to predict interactions involving temperature and river flow, two factors with weak but significant relationships with DO (0.094 for river flow). However, as shown in Table 1, its SMAPE values rose compared to TFT at both horizons, showing problems in catching abrupt short-duration changes, particularly those caused by flow variability.

Conductivity had a slight negative connection with DO (−0.15), but its effect grew more noticeable during the 30-day period. TFT and GRU successfully depicted this medium-term pattern, which is consistent with the slow impact of conductivity on river chemistry.

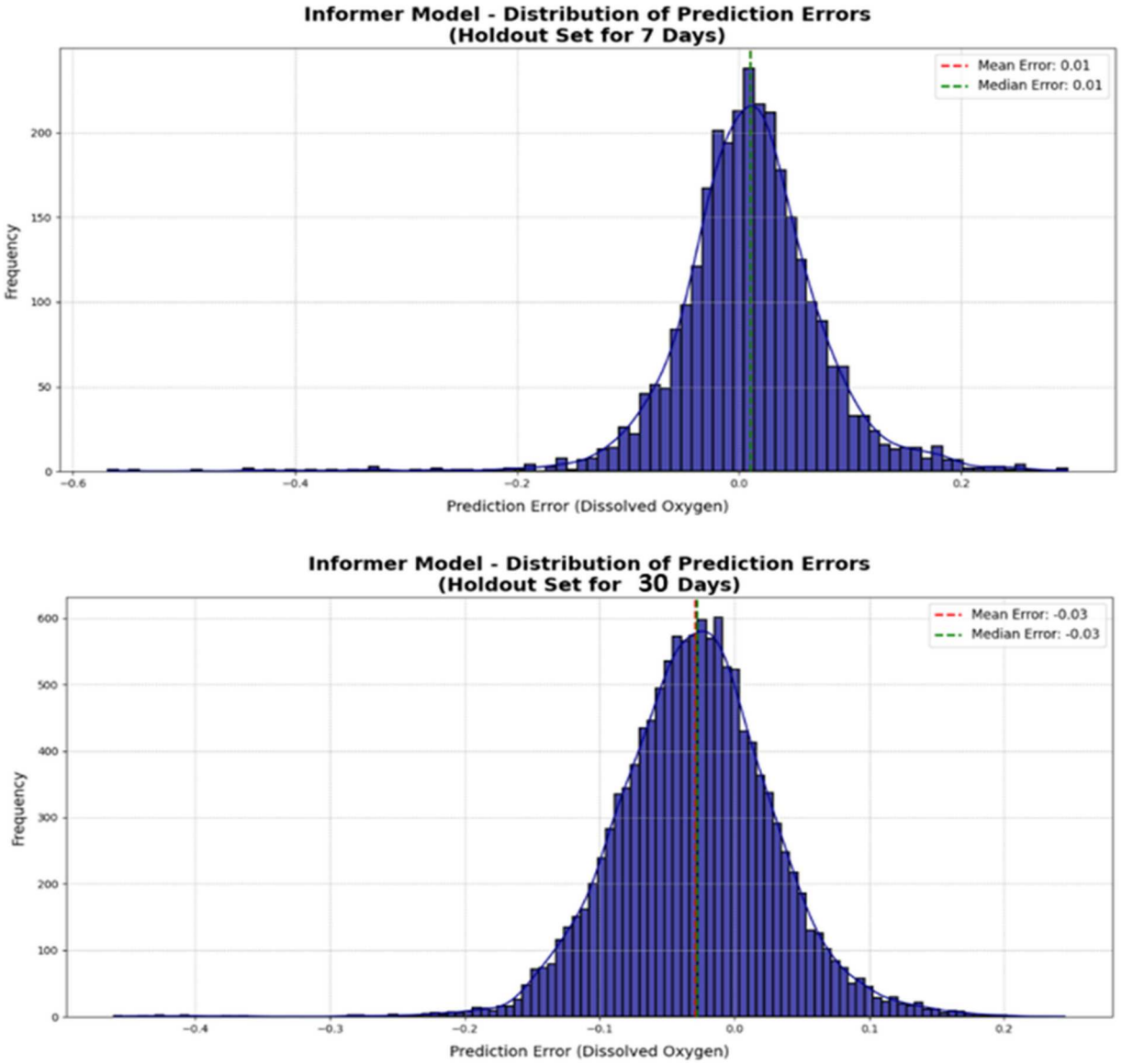

Figure 5. Informer error distribution for both horizons.

Improvements in these models' 30-day measures show that the more steady, slowly shifting character of conductivity is simpler to account for in medium-term projections. River flow had a slight positive association with DO (0.094), but was highly connected with turbidity, as indicated by their correlation of 0.53. Models sensitive to short-term hydrodynamic changes, such as TFT and LSTM, effectively captured these fluctuations after 7 days. Informer, on the other hand, struggled with sudden flow-related transitions at times, which is consistent with its architecture's preference for longer temporal relationships.

Overall, the short-term forecasting horizons revealed significant strengths in each model. TFT and LSTM were the most successful in adapting to fast hydrological changes, while GRU provided a computationally economical alternative with high immediate performance. Informer gave a balanced representation of short- and medium-range relationships. These findings highlight the necessity of choosing forecasting models that account for both the temporal properties of the target variable and the dynamic behaviour of important hydrological factors.

3.4. Premutation feature importance (PFI)

While correlation analysis is beneficial for understanding the linear correlations between hydrological factors and DO, it does not explain how these variables are used by forecasting models during prediction. To overcome this issue and improve model interpretability, a Permutation Feature Importance (PFI) analysis was performed (Figure 9). PFI is a model-independent interpretability approach that quantifies the importance of each input variable by assessing the decrease in prediction performance caused by random permutation of its values (Kaneko 2022, Khan and Byun 2023). Unlike correlation analysis, PFI takes into account nonlinear interactions and the underlying structure of deep learning

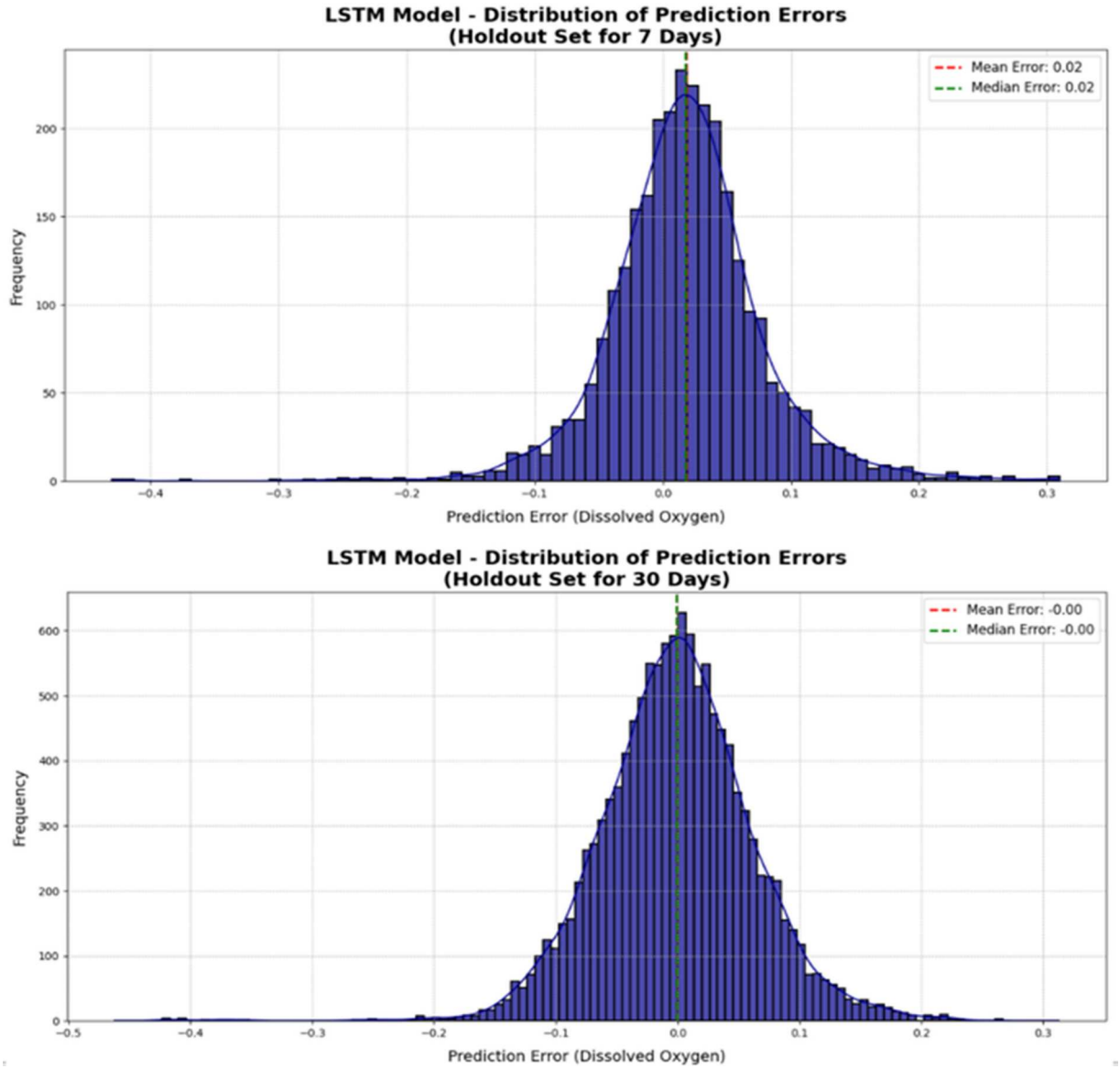

Figure 6. LSTM error distribution for both horizons.

models, making it ideal for the transformer-based and recurrent architectures employed in this work. The PFI analysis thus supplements the correlation heat map by demonstrating which hydrological drivers have the most effect on the trained forecasting model, rather than just their statistical relationship with DO.

Temperature was found as the most significant hydrological driver of DO predictors, causing the greatest rise in SMAPE when permuted (6.92). This validates temperature circumstances' dominating influence on oxygen solubility and short-term DO variability, which is compatible with well-established physical principles determining gas solubility in water. Turbidity (2.15) and conductivity (2.05) were the next most important predictors, indicating the significance of suspended particles and ionic concentration in influencing oxygen dynamics via light attenuation, mixing, and chemical reactions (Carey *et al.* 2023). Chlorophyll had a moderate impact (1.69), which is consistent with the role of algal photosynthesis and respiration in short-term DO variations. River flow (0.55) and pH (0.40) had lesser effects, showing less direct influence on DO variability at 7-30-day forecasting timeframes. These findings show that PFI gives a physically interpretable explanation for the TFT model's performance and establish that the model's major predictors are compatible with known hydrological and biogeochemical controls.

3.5. Models' limitations and strengths

The TFT, Informer, LSTM, and GRU models were tested for their strengths and limitations in forecasting dissolved oxygen (DO) across short-term horizons of 7 and 30 days. Each model revealed significant benefits based on its design, ability to capture short-term variability, and ability to represent interactions between hydrological factors. TFT performed well across both perspectives, with low SMAPE values and high R2

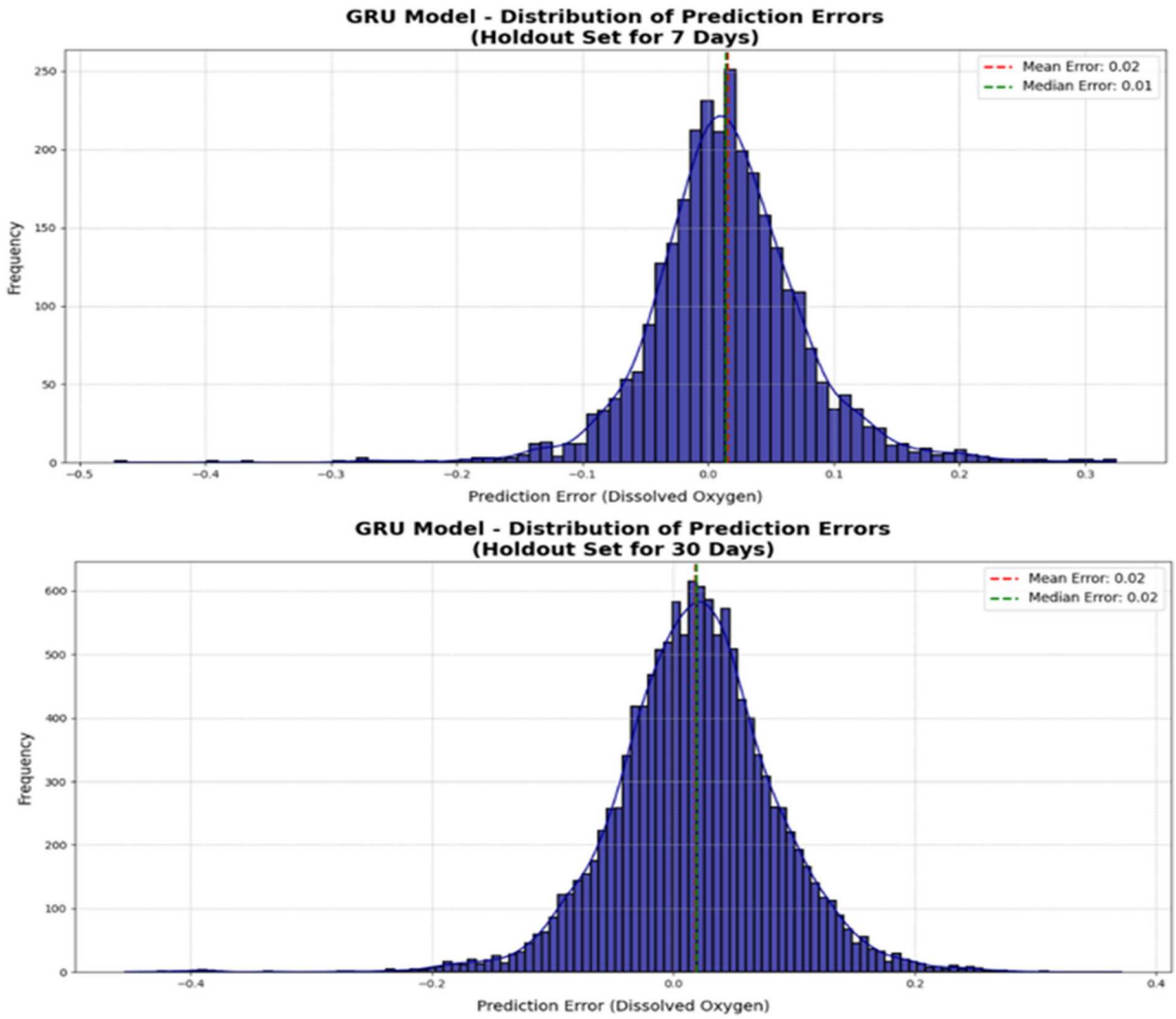

Figure 7. GRU error distribution for both horizons.

scores. Its main feature is its multi-head attention mechanism, which allows the model to prioritise informative time steps and change the impact of major factors like temperature and river flow (Lim *et al.* 2021). This ability to dynamically weight feature significance contributed to greater accuracy and narrow, centred error distributions in the findings.

LSTM also displayed significant predictive capabilities over both horizons, with performance comparable to TFT. Its sequential memory structure enabled it to successfully record short-term changes, notably those in variables like turbidity and chlorophyll (Khozani *et al.* 2022). The stability of LSTM forecasts across 7 and 30 days suggests that its design is well adapted to predicting immediate temporal relationships in hydrological data.

GRU performed similarly to LSTM, but with a simpler gating structure that allowed for effective modelling of short-term temporal patterns. Its accuracy at the 7-day horizon was comparable to more complicated structures, but its findings at 30 days revealed significantly larger error distributions, implying small decreases in stability across longer short-term windows.

Informer worked admirably on the 7- and 30-day timescales, despite being built especially for extended input sequences. Its sparse attention mechanism let it to acquire significant temporal correlations, while having somewhat higher SMAPE values than TFT. This shows that, while Informer can accurately mimic short-term dynamics, its design is less responsive to fast local fluctuations than TFT.

Overall, the comparison research shows that transformer-based models, notably TFT, have distinct benefits for short-term DO forecasting because to their capacity to collect and prioritise influential temporal aspects. Recurrent models, such as LSTM and GRU, remain great options for short-term predictions, although their stability decreases significantly when the forecasting window grows from 7 to 30 days.

3.6. Hydrological implications

The accuracy of dissolved oxygen (DO) projections is highly related to the hydrological and physicochemical features of the river system. In this investigation, short-term prediction horizons of 7 and 30 days

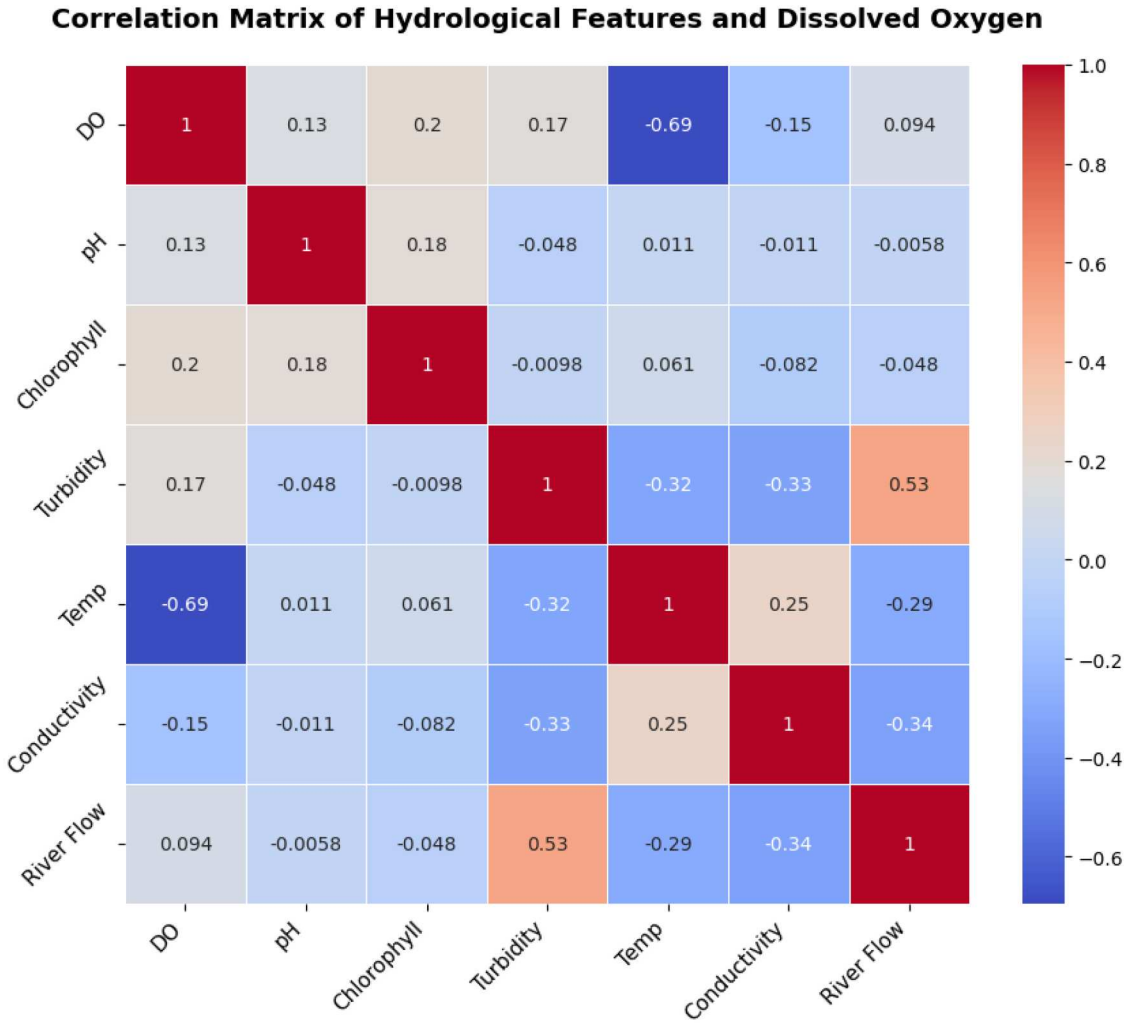

Figure 8. Heat map that describes the correlations of every feature with DO.

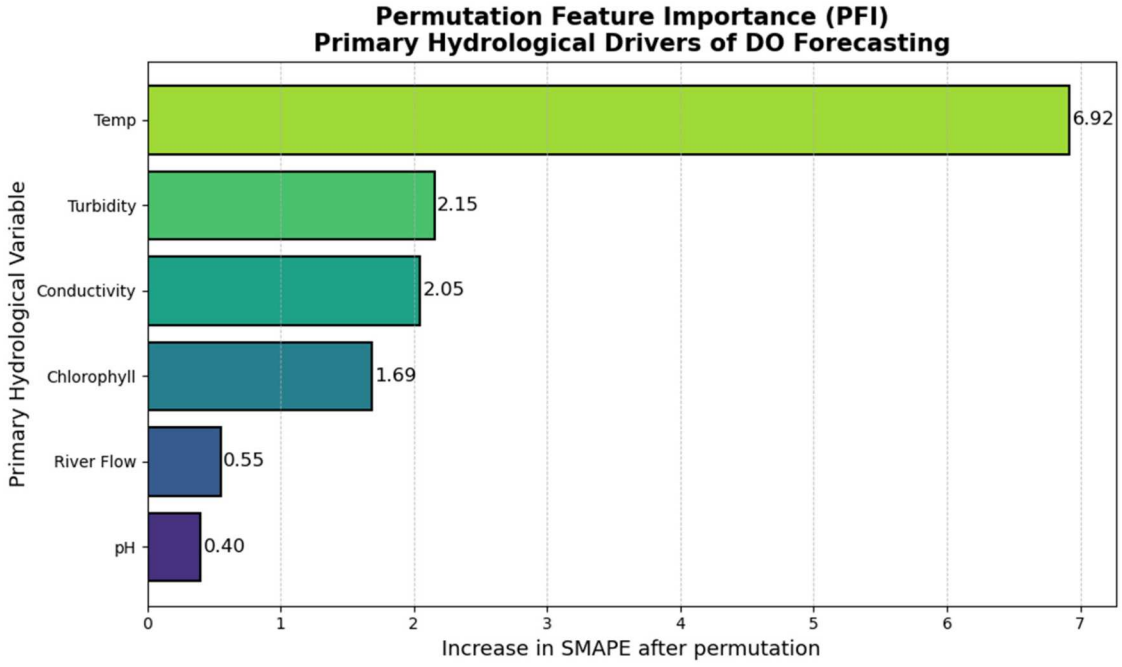

Figure 9. The permutation feature importance.

were particularly sensitive to fluctuations in temperature, turbidity, conductivity, and river flow, as evidenced by the data correlation structure and model error patterns. Temperature had the strongest link with DO, whereas turbidity and river flow were moderately related. Periods with quick changes in these determinants were linked with higher forecast errors across all models, emphasising the difficulty of capturing sudden changes in hydrological conditions within short time periods.

The TFT model's attention mechanism improved its forecast accuracy at these shorter horizons by enabling it to dynamically adjust to short-term variations in river flow and temperature. Its sensitivity to quick changes in the environment nevertheless led to minor errors, especially during times of severe weather that drastically alter water quality metrics. The Informer model balanced short-duration patterns with instantaneous fluctuations, which allowed it to function well at shorter time scales even though its design was largely focused on longer predicting horizons. Though less sensitive than TFT and LSTM, its attention processes may dynamically adapt to changing short-term hydrological conditions. Therefore, the stability and representativeness of the training data are critical factors in determining Informer's applicability at 7- and 30-day periods.

The models approached these dynamics in various ways. TFT's attention mechanism allowed it to respond to short-term fluctuations in temperature and river flow, resulting in reasonably small and centred error distributions at both horizons. LSTM and GRU, which rely on recurrent memory, caught most of the short-term temporal dependency but exhibited higher increases in error variance after 30 days, showing lower resilience as uncertainty accumulated over time. Informer provided an intermediate response: its sparse attention structure allowed it to capture both short-range and somewhat longer-range temporal patterns, but its performance lagged significantly behind that of TFT, particularly when quick local changes occurred.

These findings highlight that the efficacy of any model for short-term DO forecasting is determined not only by its design, but also by the representativeness of the training data throughout the spectrum of hydrological circumstances observed. If the training period fails to capture periods of high variability in critical factors, such as fast temperature swings or high turbidity occurrences, prediction skill will deteriorate. Careful selection and occasional update of training datasets are thus required to maintain model performance when hydroclimatic conditions change.

From a management standpoint, the findings show that models like TFT and LSTM can be useful for short-term DO prediction and operational water-quality surveillance when used within their validated time frames. Forecasts for 7–30 days can supplement normal monitoring by highlighting expected near-term conditions and assisting in the prioritisation of extra measures or research. However, they should be evaluated in conjunction with in situ observations and local expert knowledge, rather than as stand-alone decision aids, especially under situations that are beyond the range of the historical data used for training.

3.7. Limitation

Despite extensive tuning and validation, the TFT, Informer, LSTM, and GRU models have numerous limitations. First, the models were trained using data from a particular river system, so their performance reflects the site's distinct hydrological and environmental features. As a result, the findings should not be applied to other catchments without proper calibration and independent confirmation. The extra evaluation utilising data from a second monitoring station (Appendix A) lends some support to model robustness, but the shorter data record and smaller number of accessible variables restrict the strength of this evidence.

An additional issue is the models' dependence on significant historical data. While deep learning algorithms may capture complicated temporal patterns, they require sufficiently large and continuous datasets, which may not be available in areas with limited monitoring infrastructure. This limitation limits the models' direct transferability to data-poor systems and emphasises the significance of constant long-term environmental monitoring for successful predictive prediction.

Finally, while tactics like validation splits, early halting, and regularisation were used to prevent overfitting, the risk of model over-specialisation to the training dataset cannot be completely ruled out. Broader testing over more rivers and under a wider variety of hydrological circumstances would improve knowledge of generalisability and aid in identifying scenarios where model performance may decrease. Future research should incorporate multi-site evaluations and longer temporal records to test the predictability of these forecasting systems across a variety of environmental conditions.

4. Summary and conclusions

This study compared the performance of four deep learning models – Temporal Fusion Transformer (TFT), Informer, Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) – for forecasting dissolved oxygen (DO) in a river system over 7 and 30 days. The investigation revealed that, while each model identified fundamental hydrological

linkages, their forecasting accuracy varied depending on design and capacity to describe short-range temporal correlations. The ability of the Temporal Fusion Transformer (TFT), Informer, Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) models to predict dissolved oxygen (DO) values across short time horizons of 7- and 30-days was thoroughly evaluated in this study. Due in large part to different designs and capacity to manage abrupt hydrological variations and environmental unpredictability, each model showed unique advantages and disadvantages.

Temperature had the highest negative connection with DO, demonstrating its central role in short-term oxygen dynamics. Turbidity and river flow also had a significant impact on model behaviour, indicating their influence on hydrological variability. Adding delayed and rolling-window features increased model performance by capturing recent environmental changes and smoothing out short-term swings. time monitoring systems to support sustainable water quality management.

Overall, the work emphasises the advantages of attention-based architectures, notably TFT, for short-term DO forecasting while demonstrating the ongoing significance of recurrent models like as LSTM and GRU over shorter time horizons. These findings give practical assistance for selecting forecasting models appropriate for different temporal scales in river water-quality management. While the models have a high potential for short-term decision-making, their use should be supplemented with continuing monitoring and site-specific calibration to ensure dependability throughout different hydrological conditions.

Disclosure statement

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

Funding

This study is partially funded by the UK Research and Innovation UKRI project 10063665.

ORCID

Ali J. Ali http://orcid.org/0009-0007-8359-7787

References

Ahmed, A.A., *et al.*, 2024. Applications of machine learning to water resources management: a review of present status and future opportunities. *Journal of Cleaner Production*, 441, 140715.

Ali, A.J. and Ahmed, A.A., 2024. Long-term AI prediction of ammonium levels in rivers using transformer and ensemble models. *Cleaner Water*, 2, 100051.

Ali, A. J. and Ahmed, A. A., 2025. Aquifer-specific flood forecasting using machine learning: a comparative analysis for three distinct sedimentary aquifers. *Science of The Total Environment*, 1004, 180756.

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*, 484, 144300.

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

Cahuantzi, R., Chen, X., and Güttel, S., 2023. A comparison of LSTM and GRU networks for learning symbolic sequences . *In*: *Science and information conference*. Cham: Springer Nature Switzerland, 771–785.

Carey, C.C., Lewis, A.S., and Breef-Pilz, A., 2023. Time series of high-frequency profiles of depth, temperature, dissolved oxygen, conductivity, specific conductance, chlorophyll a, turbidity, pH, oxidation-reduction potential, photosynthetic active radiation, and descent rate for Beaverdam Reservoir, Carvins Cove Reservoir, Falling Creek Reservoir, Gatewood Reservoir, and Spring Hollow Reservoir in southwestern Virginia, USA 2013-2022.

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

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

Choi, S.R. and Lee, M., 2023. Transformer architecture and attention mechanisms in genome data analysis: a comprehensive review. *Biology*, 12 (7), 1033.

Chung, J., *et al.*, 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

Clevert, D.A., Unterthiner, T., and Hochreiter, S., 2015. Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:1511.07289.

Cox, B.A., 2003. A review of dissolved oxygen modelling techniques for lowland rivers. *The Science of The Total Environment*, 314-316, 303–334.

Danladi Bello, A.A., Hashim, N.B., and Mohd Haniffah, M.R., 2017. Predicting impact of climate change on water temperature and dissolved oxygen in tropical rivers. *Climate*, 5 (3), 58.

Dayal, A., *et al.*, 2024. Deep learning for multi-horizon water levelForecasting in KRS reservoir, India. *Results in Engineering*, 21, 101828.

Deepa, B. and Ramesh, K., 2022. Epileptic seizure detection using deep learning through min max scaler normalization. *International Journal of Health Sciences*, 6, 10981–10996.

Deshpande, P., *et al.*, 2021, March. Long horizon forecasting with temporal point processes. *In: Proceedings of the 14th ACM international conference on web search and data mining*, 571–579.

Dey, R. and Salem, F.M., 2017, August. Gate-variants of gated recurrent unit (GRU) neural networks. *In: 2017 IEEE 60th international midwest symposium on circuits and systems (MWSCAS)*, IEEE, 1597–1600.

El-Nahhal, D., *et al.*, 2021. Acidity, electric conductivity, dissolved oxygen total dissolved solid and salinity profiles of

marine water in Gaza: influence of wastewater discharge. *American Journal of Analytical Chemistry*, 12 (11), 408–428.

Ghobadi, F., Yaseen, Z.M., and Kang, D., 2024. Long-term streamflow forecasting in data-scarce regions: insightful investigation for leveraging satellite-derived data, informer architecture, and concurrent fine-tuning transfer learning. *Journal of Hydrology*, 631, 130772.

Gundu, V. and Simon, S.P., 2021. PSO–LSTM for short term forecast of heterogeneous time series electricity price signals. *Journal of Ambient Intelligence and Humanized Computing*, 12, 2375–2385.

Haider, H., Ali, W., and Haydar, S., 2013. Evaluation of various relationships of reaeration rate coefficient for modeling dissolved oxygen in a river with extreme flow variations in Pakistan. *Hydrological Processes*, 27 (26), 3949–3963.

Huey, G.M. and Meyer, M.L., 2010. Turbidity as an indicator of water quality in diverse watersheds of the Upper Pecos River Basin. *Water*, 2 (2), 273–284.

Irvine, K.N., *et al.*, 2011. Spatial and temporal variability of turbidity, dissolved oxygen, conductivity, temperature, and fluorescence in the lower Mekong River–Tonle Sap system identified using continuous monitoring. *International Journal of River Basin Management*, 9 (2), 151–168.

Kaneko, H., 2022. Cross-validated permutation feature importance considering correlation between features. *Analytical Science Advances*, 3 (9-10), 278–287.

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.

Khan, P.W. and Byun, Y.C., 2023. Optimized dissolved oxygen prediction using genetic algorithm and bagging ensemble learning for smart fish farm. *IEEE Sensors Journal*, 23 (13), 15153–15164.

Khozani, Z.S., *et al.*, 2022. Combining autoregressive integrated moving average with long short-term memory neural network and optimisation algorithms for predicting ground water level. *Journal of Cleaner Production*, 348, 131224.

Kim, Y.W., *et al.*, 2021. Forecasting abrupt depletion of dissolved oxygen in urban streams using discontinuously measured hourly time-series data. *Water Resources Research*, 57 (4), e2020WR029188.

Kney, A.D. and Brandes, D., 2007. A graphical screening method for assessing stream water quality using specific conductivity and alkalinity data. *Journal of Environmental Management*, 82 (4), 519–528.

Kong, T., *et al.*, 2021. Computer vision and long short-term memory: learning to predict unsafe behaviour in construction. *Advanced Engineering Informatics*, 50, 101400.

Kulkarni, S.J., 2016. A review on research and studies on dissolved oxygen and its affecting parameters. *International Journal of Research and Review*, 3 (8), 18–22.

Kushwaha, N.L., *et al.*, 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*, *10*(10), e31085.

Lamping, J., *et al.*, 2005. Effectiveness of aeration and mixing in the remediation of a saline stratified river. *Environmental Science & Technology*, 39 (18), 7269–7278.

Lessels, J.S. and Bishop, T.F.A., 2020. A post-event stratified random sampling scheme for monitoring event-based water quality using an automatic sampler. *Journal of Hydrology*, 580, 123393.

Li, D., *et al.*, 2025. A long-term dissolved oxygen prediction model in aquaculture using transformer with a dynamic adaptive mechanism. *Expert Systems with Applications*, 259, 125258.

Li, S., *et al.*, 2024. Explainable machine learning models for estimating daily dissolved oxygen concentration of the tualatin river. *Engineering Applications of Computational Fluid Mechanics*, 18 (1), 2304094.

Lim, B., *et al.*, 2021. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37 (4), 1748–1764.

Luo, A., *et al.*, 2024. The impact of rainfall events on dissolved oxygen concentrations in a subtropical urban reservoir. *Environmental Research*, 244, 117856.

Maldonado-Cruz, E. and Pyrcz, M.J., 2024. Multi-horizon well performance forecasting with temporal fusion transformers. *Results in Engineering*, 21, 101776.

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

Martín-Suazo, S., *et al.*, 2024. Deep learning methods for multi-horizon long-term forecasting of harmful algal blooms. *Knowledge-Based Systems*, 301, 112279.

Mertikas, S., 2023. Error distribution and accuracy measure in navigation: An overview.

Neal, C., *et al.*, 2006. Chlorophyll-a in the rivers of eastern England. *Science of the Total Environment*, 365 (1-3), 84–104.

Pant, N., Toshniwal, D., and Gurjar, B.R., 2024. Multi-step forecasting of dissolved oxygen in River Ganga based on CEEMDAN-AdaBoost-BiLSTM-LSTM model. *Scientific Reports*, 14 (1), 11199.

Pena, M.A., *et al.*, 2010. Modeling dissolved oxygen dynamics and hypoxia. *Biogeosciences*, 7 (3), 933–957.

Quaedvlieg, R., 2021. Multi-horizon forecast comparison. *Journal of Business & Economic Statistics*, 39 (1), 40–53.

Radwan, M., *et al.*, 2003. Modelling of dissolved oxygen and biochemical oxygen demand in river water using a detailed and a simplified model. *International Journal of River Basin Management*, 1 (2), 97–103.

Rajesh, M. and Rehana, S., 2022. Impact of climate change on river water temperature and dissolved oxygen: Indian riverine thermal regimes. *Scientific Reports*, 12 (1), 9222.

Sseguya, F. and Jun, K.S., 2024. Deep learning ensemble for flood probability analysis. *Water*, 16 (21), 3092.

Vafaei, N., Ribeiro, R.A., and Camarinha-Matos, L.M., 2018. Data normalisation techniques in decision making: case study with TOPSIS method. *International Journal of Information and Decision Sciences*, 10 (1), 19–38.

Vaswani, A., *et al.*, 2017. Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

Wang, C. and Tang, W., 2023. Temporal fusion transformer-Gaussian process for multi-horizon river level prediction and uncertainty quantification. *Journal of Circuits, Systems and Computers*, 32 (18), 2350309.

Wang, N. and Zhao, X., 2023. Enformer: Encoder-based sparse periodic self-attention time-series forecasting. *IEEE Access*, 11, 112004–112014.

Xiao, X., *et al.*, 2023. The efficiency of the microbial carbon pump as seen from the relationship between apparent oxygen utilization and fluorescent dissolved organic matter. *Progress in Oceanography*, 210, 102929.

Xu, C., Chen, X., and Zhang, L., 2021. Predicting river dissolved oxygen time series based on stand-alone models and hybrid wavelet-based models. *Journal of Environmental Management*, 295, 113085.

Xu, H., *et al.*, 2023. Power-load forecasting model based on informer and its application. *Energies*, 16 (7), 3086.
Zhi, W., *et al.*, 2023. Temperature outweighs light and flow as the predominant driver of dissolved oxygen in US rivers. *Nature Water*, 1 (3), 249–260.
Zhou, H., *et al.*, 2021, May. Informer: beyond efficient transformer for long sequence time-series forecasting. *In:* Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 12, pp. 11106–11115).

Appendix

The results from a second monitoring station are including in this section, which was added to confirm the accuracy and generalisability of our forecasting models in various geological contexts. Despite having a smaller scope, the data from this station provides insightful information about how well the models function in a particular environmental setting. Tables A1 and A2 present the validation and the holdout results across both horizons.

Table A1. The validation results of all models across all 4 horizons for Station 2.

	7-Step Ahead (Validation)			
Model	RMSE	MAE	R^2	SMAPE
TFT	0.09	0.07	0.74	**11.06**
Informer	0.09	0.06	0.74	**10.73**
LSTM	0.08	0.06	0.78	**9.87**
GRU	0.07	0.05	0.75	**9.03**

Model	30-Step Ahead (Validation)			
	RMSE	MAE	R^2	SMAPE
TFT	0.08	0.06	0.76	11.58
Informer	0.09	0.07	0.73	12.39
LSTM	0.09	0.06	0.74	11.73
GRU	0.08	0.06	0.65	11.13

Table A2. The holdout results of all models across all 4 horizons for Station 2.

	7-Step Ahead (Holdout)			
Model	RMSE	MAE	R^2	SMAPE
TFT	0.08	0.06	0.81	**10.74**
Informer	0.08	0.06	0.81	**11.00**
LSTM	0.07	0.06	0.83	**10.21**
GRU	0.06	0.04	0.79	**8.47**

	30-Step Ahead (Holdout)			
Model	RMSE	MAE	R^2	SMAPE
TFT	0.08	0.06	0.79	11.13
Informer	0.08	0.06	0.77	11.79
LSTM	0.08	0.06	0.77	11.47
GRU	0.07	0.05	0.70	9.76

Compared to the major datasets discussed in the main text, the dataset from the second monitoring station has fewer variables including (conductivity, temperature and turbidity), in addition, it covers a shorter time span. Despite these drawback, the research offers helpful information about the flexibility and efficacy of the models.

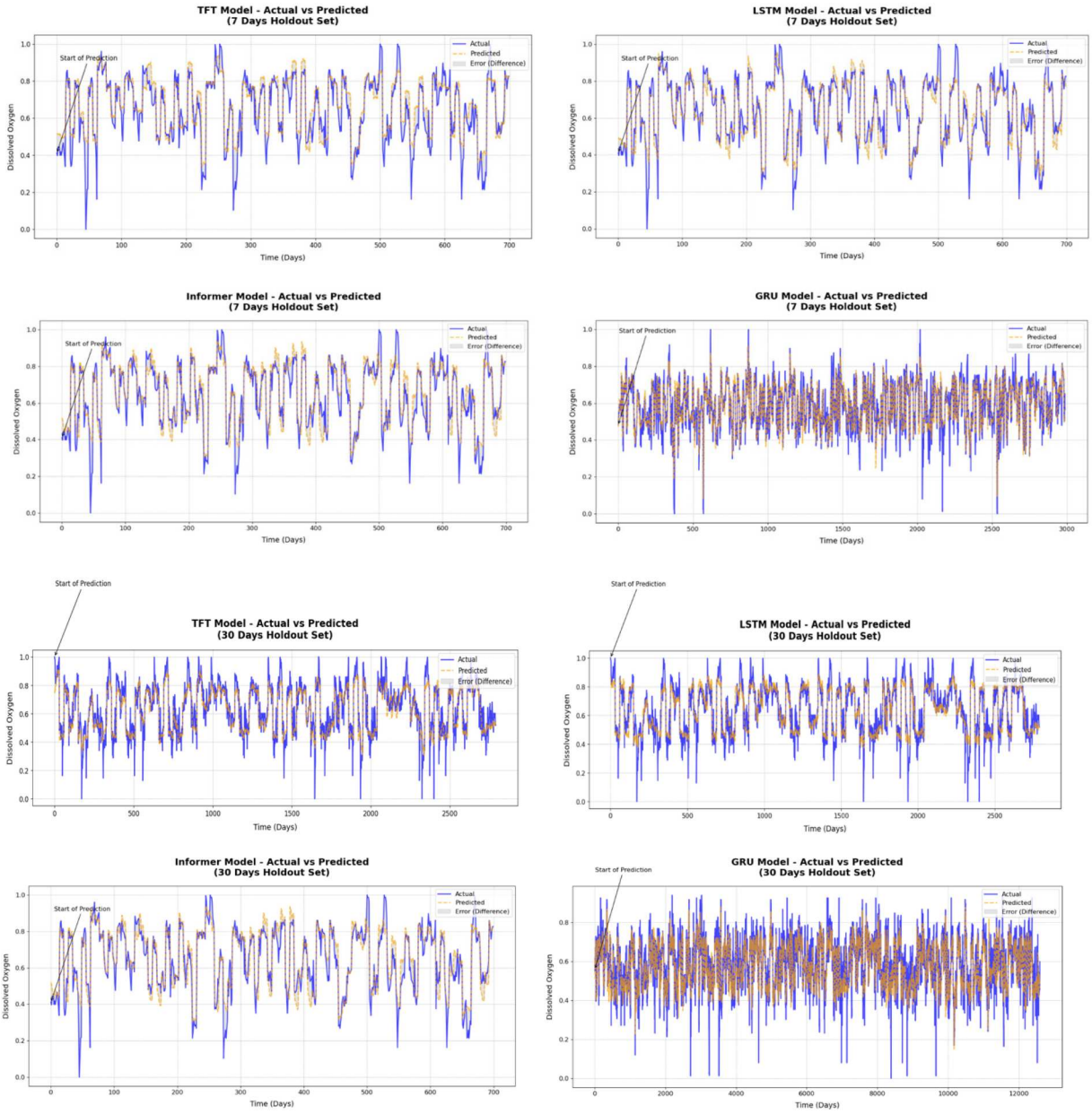

Figure A1. Actual vs predicted for short-term hoirizon.

The superiority of transformer-based models (TFT and Informer) over classic recurrent models (LSTM and GRU) in managing short- and mid-term forecasts with high accuracy is demonstrated by the comparison of these models over these time horizons. For applications like water quality management and environmental monitoring that demand accurate and trustworthy forecasts, this distinction is essential. For situations where quick, real-time DO level estimates are crucial, the recurrent models continue to provide useful capabilities despite their limits for longer time horizons. These observations not only support the reliability of the models selected for this investigation, but they also direct future advancements and applications in environmental sciences predictive modelling.

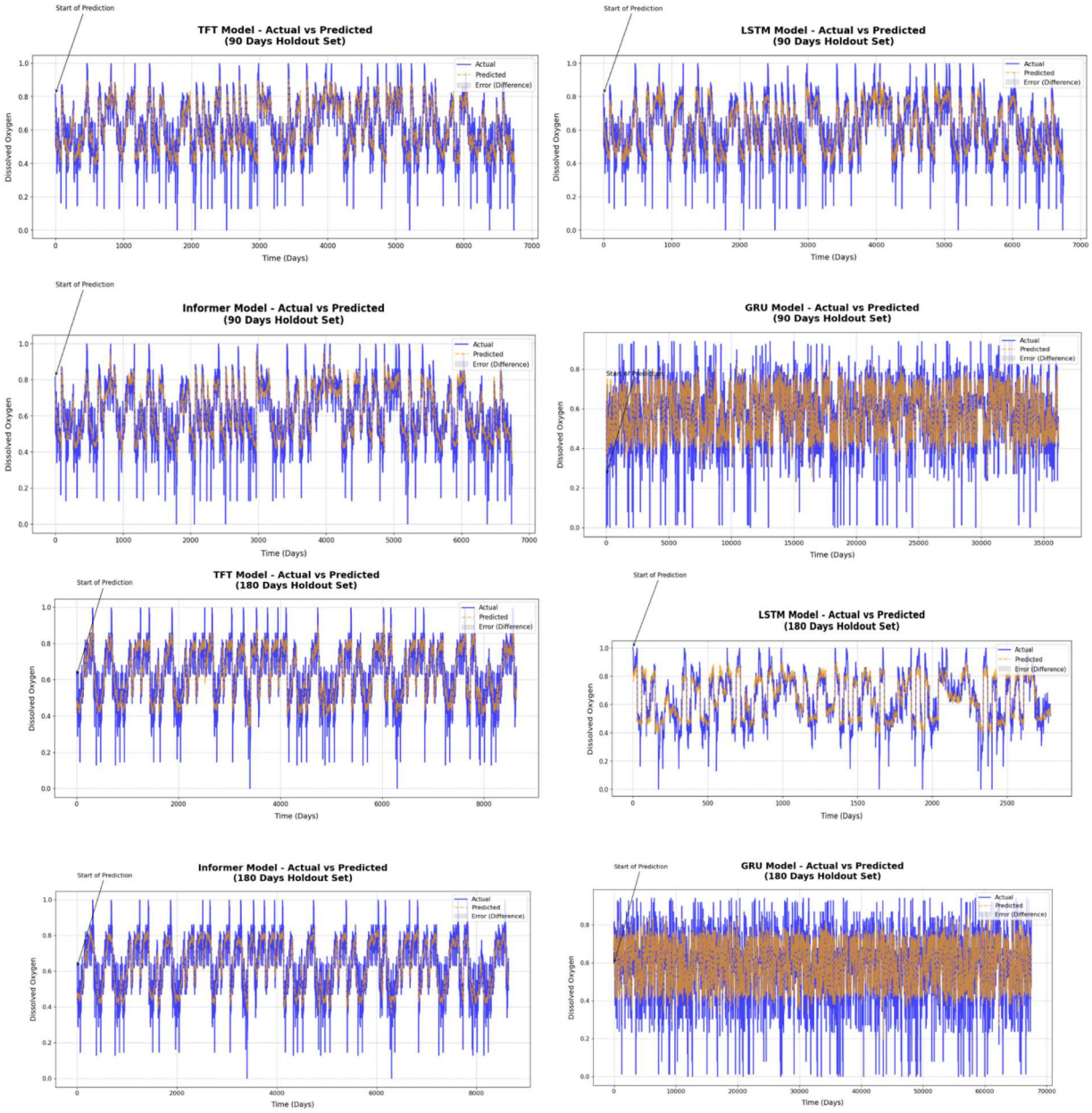

Figure A2. Actual vs predicted for short-term hoirizon.

Figure A2 shows how the performance of each model changes as the forecast horizon increases. They show that although all models are competent in the short term, transformer-based models – especially the Informer – consistently perform better as the forecast period lengthens. For those involved in environmental management, these insights are essential since the precision of long-term projections has a big influence on policymaking and decision-making. By providing a nuanced view of each model's capabilities across various time frames, incorporating this thorough study into the text will improve the findings section. It also offers helpful advice for choosing the right models according to the particular requirements of environmental monitoring programmes, making sure that the models selected match the time scales of the ecological processes they are meant to forecast.