Science of the Total Environment 1004 (2025) 180756

Contents lists available at ScienceDirect

Science of the Total Environment

journal homepage: www.elsevier.com/locate/scitotenv

Aquifer-specific flood forecasting using machine learning: A comparative analysis for three distinct sedimentary aquifers

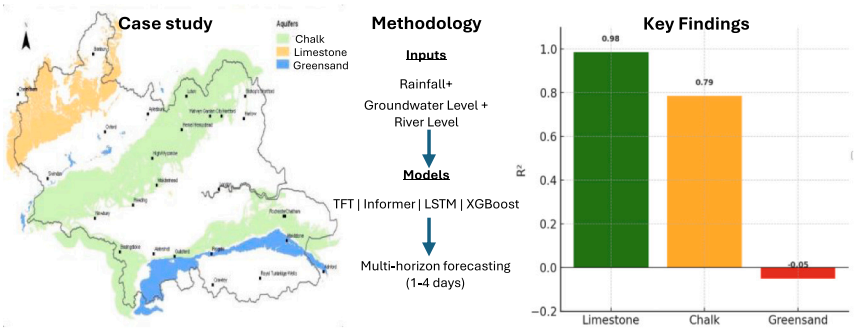
Ali J. Ali [a,*], Ashraf A. Ahmed [a,*]

[a] Department of Civil and Environmental Engineering, Brunel University London, Uxbridge, UB8 3PH, United Kingdom

HIGHLIGHTS

- Transformer models outperform XGBoost across Chalk and Limestone aquifers.
- Continuous flood forecasting captures dynamics better than binary thresholds.
- Limestone aquifer predictions achieve very high accuracy ($R^2 = 0.98–0.99$).
- Chalk aquifer models show moderate accuracy ($R^2 \approx 0.77–0.80$).
- Models in Greensand perform poorly ($R^2 \leq 0$, negative predictability).

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Deep learning
Flood risk management
Groundwater–river interaction
Temporal fusion transformer
Multi-horizon forecasting

ABSTRACT

Accurate flood prediction is critical for avoiding catastrophic impacts, but its difficulty varies by geological location. This study evaluates four machine learning models – TFT, Informer, LSTM, and XGBoost – for multi-horizon flood forecasting (1-4 days), across Limestone, Chalk, and Greensand located in the Thames Basin, UK. Stations were carefully chosen using the UK government flood risk maps, geological mapping, and Environment Agency hydrological data to guarantee a complete portrayal of aquifer-specific groundwater-river interactions. The results show that the model accuracy varies significantly depending on aquifer features. Rapid GWL-river interactions allowed Limestone aquifers to achieve very high precision (R^2 = 0.98–0.99), with transformers and LSTM clearly surpassing XGBoost. The accuracy of Chalk aquifers was moderate (R^2 = 0.77–0.80), indicating delayed reactions and intermediate permeability. Greensand aquifers were difficult to model due to delayed and complex reactions, resulting in low or negative R^2 values. Correlation study confirmed these findings: Limestone showed a significant groundwater-river linkage ($r = 0.84$), Chalk moderate ($r = 0.26$), and Greensand had a small negative association ($r = -0.14$). The novelty of this study highlights the significant impact of subsurface hydrology on predicted reliability, revealing aquifer-specific geological restrictions in ML-based forecasting. This research offers a more physically consistent early warning method by fusing GWL data with developed transformer architectures. The results highlight the significance of adjusting forecasting frameworks to geological environments, which has direct implications for resilience planning and flood risk management at the watershed scale.

* 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.scitotenv.2025.180756
Received 24 June 2025; Received in revised form 15 October 2025; Accepted 15 October 2025

1. Introduction

Floods are one of the most frequent and destructive natural catastrophes in the world, and reducing their negative effect requires accurate flood forecasts (Tellman et al., 2021). Accurate forecasting helps with long-term planning, infrastructure design, and flood-defence tactics in addition to emergency reaction and evacuation (Brunner et al., 2021; Hamel and Tan, 2022). Advanced forecasting techniques are becoming increasingly important as climate change intensifies extreme rains, raises sea levels, and changes hydrological regimes (Tabari, 2020; Ahmed et al., 2024). The intricate relationships between land surfaces, human activity, and climatic variables make flood prediction fundamentally difficult. The geological differences between Chalk, Limestone, and Greensand aquifers result in diverse groundwater-river reactions, complicating predictions (Perzan et al., 2023).

Predicting and managing floods is particularly difficult in the Thames Basin, which includes both rural areas like Oxfordshire and Gloucestershire and urban areas like London (Bearcock and Smedley, 2010). These issues are the result of a combination of varied landscapes, climate change, and human activity. Urbanisation in Greater London has disrupted natural drainage, increased surface runoff and decreased soil permeability (Jenkins et al., 2018; Environment Agency, 2022). Tidal, surface water, groundwater, and fluvial floods all affect the Basin, necessitating different mitigation and forecasting strategies (Environment Agency, 2022).

Climate change intensifies threats by increasing the frequency of extreme weather events and rising sea levels, specifically in the lower Thames. Large catchments also respond slowly to rainfall, necessitating the use of hydrological and meteorological models for forecasting (Crooks and Kay, 2015; Ali et al., 2024). Effective long-term management necessitates cross-agency coordination within complicated regulatory frameworks (Fan, 2024), as well as advanced forecasting techniques customised to basin conditions.

Recent advancements integrate hydrological and meteorological measurements with real-time monitoring and machine learning (ML). Distributed and physically based models can capture regional variability, but their processing requirements restrict their use (Hussain et al., 2021; Wang et al., 2023). To address these restrictions, data-driven solutions have gained popularity. Surrogate modelling provides efficient approximations of complicated flood processes (Donnelly et al., 2022), but physics-informed neural networks include hydrological restrictions directly into training, enhancing physical consistency (Donnelly et al., 2024).

With the increased availability of large data from ground sensors, radar, and satellites, AI-driven techniques have shown higher accuracy and lead times, which are critical for early warning and resource allocation (Motta et al., 2021; Yuan et al., 2022). These developments represent a significant move towards integrated, adaptive flood forecasting that uses ML to overcome the challenges of solely empirical or physics-based models.

Transformer-based models, such as the Temporal Fusion Transformer (TFT), use gating and attention processes to capture both short- and long-term relationships whilst providing interpretability for hydrological data (Lim et al., 2021; Ali and Ahmed, 2024). The Informer improves on these techniques by including a ProbSparse self-attention mechanism, making it more efficient for lengthy sequences and appropriate for longer flood predicting horizons (Zhou et al., 2021; Wang and Zhao, 2023).

Extreme Gradient Boosting (XGBoost) is still a popular ensemble approach, appreciated for its scalability, tolerance to missing information, and good performance in rainfall-runoff and flood risk applications (Chen and Guestrin, 2016). Long Short-term Memory (LSTM), as recurrent models, are still used to forecast river flow, rainfall, and flood events due to their ability to learn long-term dependencies (Hochreiter and Schmidhuber, 1997; Noh, 2021).

This research uses these four models to present a comparative framework for aquifer-specific, multi-horizon flood forecasting. By comparing transformer topologies to conventional and recurrent approaches, we can see how model design and geological context impact forecast reliability in the Thames Basin. By combining hydrological records of rainfall, groundwater levels, and river stages with cutting-edge machine learning algorithms, this project seeks to enhance flood forecasting and risk management in the Thames Basin. In order to ascertain how aquifer-specific variables affect prediction reliability, we specifically assess the performance of four top models – Temporal Fusion Transformer (TFT), Informer, Long Short-Term Memory (LSTM), and XGBoost – across short-term horizons (1–4 days).

This study offers the first aquifer-specific, multi-horizon comparative analysis of flood forecasting in the Thames Basin, combining cutting-edge machine learning with geological context. Unlike prior research that assumed catchments were hydrologically homogeneous, our approach explicitly shows how geological variations across Chalk, Limestone, and Greensand aquifers affect forecast dependability. This is the first demonstration that transformer-based models may detect consistent differences in predictability caused by subsurface controls. The peculiarity of this work is that it not only benchmarks sophisticated models but also links their performance to aquifer features, demonstrating that even cutting-edge designs might fail if subsurface dynamics are disregarded. By creating this link, we introduce a new paradigm for flood forecasting that is both data-driven and hydrogeologically informed. This contribution has immediate implications for operational agencies like the Environment Agency, which can use aquifer-aware forecasting to provide more reliable, region-specific flood warnings, as well as researchers looking for reproducible, physically consistent methods of incorporating groundwater-river interactions into machine learning-based forecasting.

2. Methodology

2.1. Study area

2.1.1. Thames Basin overview

The Thames Basin is one of the biggest and most hydrologically varied areas in the United Kingdom. From northern Oxfordshire and Gloucestershire to the Thames Estuary and portions of Kent, including the heavily populated metropolitan area of Greater London, it covers an area of more than 16,200 km^2 (Bearcock and Smedley, 2010). Wide floodplains, slow-responding river systems, and considerable tidal effect in the lower reaches contribute to a complicated flood-risk profile. Floods have already damaged millions in the watershed, necessitating ongoing monitoring and numerous risk-management techniques (Environment Agency, 2022). Ali et al. (2024) have highlighted the necessity for forecasting models in this area by demonstrating the crucial role that groundwater dynamics play in flood prediction.

2.1.2. Station selection

The choice of stations for this study was based on a deliberate focus on high-risk flood locations as determined by the UK's long-term flood risk service maps to capture a variety of hydrological settings throughout the Thames Basin Environment Agency (2022). The monitoring sites were selected to symbolise the three main aquifer types in the area, as shown in Fig. 1: Greensand, Limestone and Chalk.

Chalk: High secondary porosity due to fracture networks, significant permeability, and often delayed groundwater reactions (Smedley et al., 2003; Shand et al., 2003a; Neal et al., 2006; Environment Agency, 2022).

Limestone: Jurassic limestone, with confined/ unconfined settings and fracture-controlled flow, can provide more dynamic groundwater reactions during rainfall events (Oubagaranadin et al., 2007).

Greensand: The greensand aquifer, which is usually found next to Chalk formations, has special geological features that provide exceptional transmissivity and storage qualities (Shand et al., 2003b).

To ensure comparability, three monitoring sites were selected within each aquifer type that measure rainfall, groundwater levels, and river levels simultaneously. Site selection was based on (i) geographic closeness of measurements, (ii) high flood risk sites, and (iii) uniform temporal coverage. Data availability extended from April 2011 to early 2025, with modest variations by aquifer. Figs. 1 and 2 depict the Thames Basin overview, aquifer borders and flood-risk zones.

2.1.3. Data sources and pre-processing

The Environment Agency's Hydrological Data Explorer was largely used to access hydrological datasets such as rainfall, groundwater levels (GWL), and river levels, with local weather stations providing support. Hourly observations were aggregated to daily averages to guarantee uniformity across variables and reduce the impact of infrequent missing data (Kang and Tian, 2018; Environment Agency, 2022) (Fig. 3).

The remaining gaps were filled using linear interpolation, a commonly used approach for hydrological time series that maintains temporal continuity while reducing biases in model training and assessment (Kang and Tian, 2018). To facilitate feature comparability and increase model stability, all variables were normalised to a 0–1 scale using the MinMaxScaler method (Deepa and Ramesh, 2022). To capture delayed hydrological responses and decrease noise in short-term fluctuations, lagged features (1–3 days) and 3-day rolling averages for rainfall, river level, and GWL were created. These designed predictors improve the depiction of aquifer-specific dynamics, where temporal delays and storage effects play a significant role in flooding processes.

This structured preprocessing methodology ensures that data is consistent and comparable among aquifers, which is essential for constructing credible flood forecasting models. Supplementary Information (S2) includes detailed pretreatment techniques and implantation parameters.

2.2. Model development

Four sophisticated ML models were used to assess aquifer-specific flood forecasting ability over several time periods (1–4). These comprise two transformer-based architectures, a recurrent neural network, and a gradient-boosting baseline. They present a balanced evaluation of sequence learning, attention-based forecasting and ensemble techniques.

2.2.1. Long short-term memory

The LSTM network is a recurrent neural network that is intended to solve the vanishing gradient problem found in ordinary RNNs. Its gated structure (input, forget, and output gates) controls information flow, allowing for the learning of both short- and long-term dependencies (Hochreiter and Schmidhuber, 1997; Khozani et al., 2022). This is crucial in hydrology since floods are caused by both rapid rainfall and delayed groundwater contribution (Le et al., 2019). LSTM is commonly used in rainfall-runoff modelling, groundwater level forecasting, and streamflow prediction (Kratzert et al., 2019; Shen, 2018; Ali et al., 2024). In this work, LSTM serves as a benchmark deep learning model for capturing temporal memory in aquifer river interactions. The Supplementary Information (S3) contains details of gate equations, optimiser settings, and training parameters.

2.2.2. Temporal Fusion Transformer (TFT)

The TFT is a hybrid architecture that combines recurrent layers with multi-head attention and gating methods to provide interpretable multi-horizon forecasting (Lim et al., 2021). Unlike LSTM, TFT may dynamically weight inputs using variable selection networks and attention heads, delivering information on feature relevance over time. This is especially useful for aquifer-specific flood forecasting because the effects of rainfall, river levels, and groundwater might vary over time (Ali et al., 2024). Because of its capacity to simulate nonlinear, non-stationary connections, the TFT has performed well in hydrological and environmental prediction tasks (Koya and Roy, 2024). Supplementary Information (S3) contains mathematical formulations of both the attention mechanism and gated residual networks.

2.2.3. Informer

The informer is a transformer-based model designed for efficient

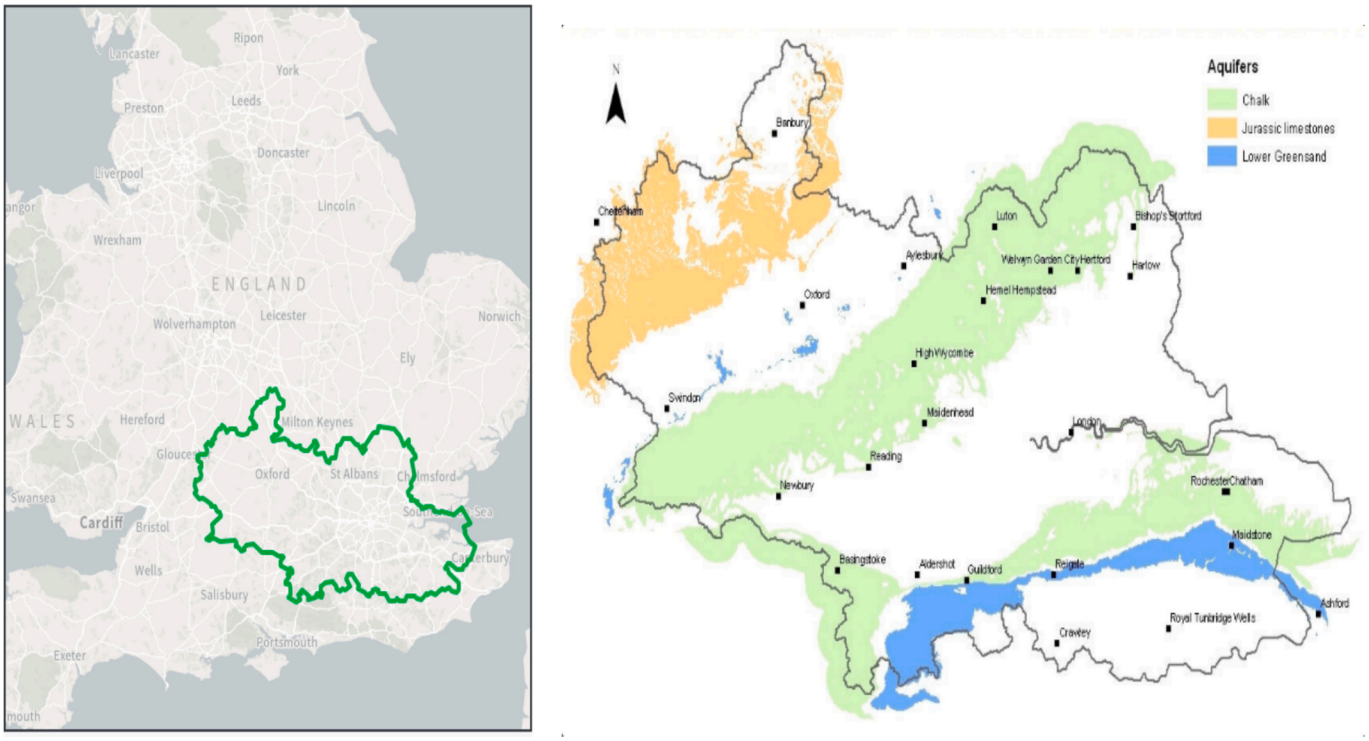
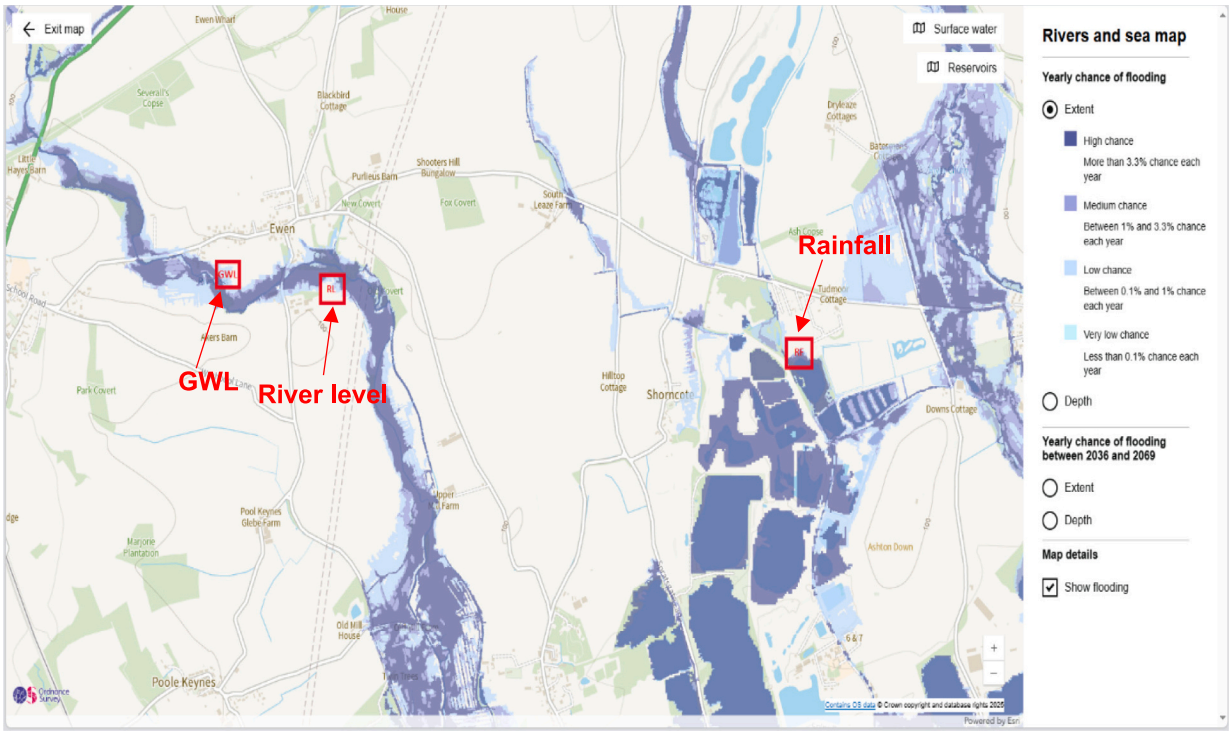


Fig. 1. Thames Basin (Ali et al., 2024).

i)



ii)

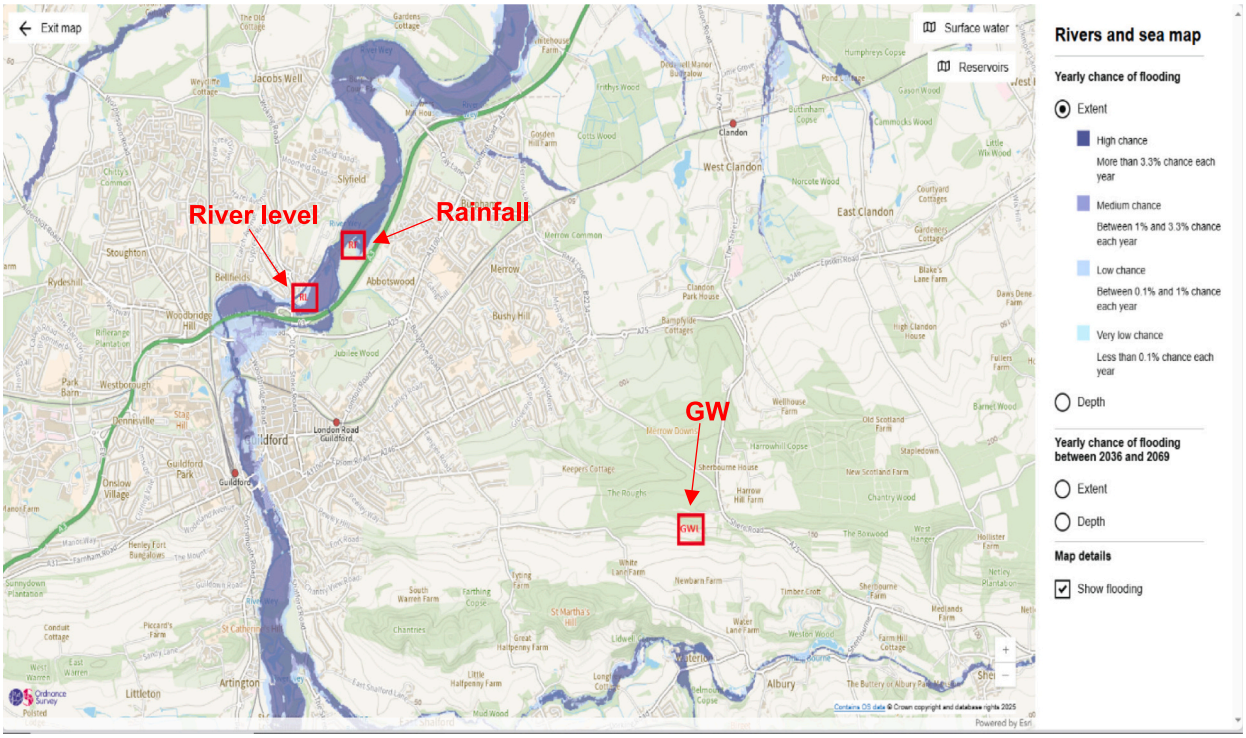


Fig. 2. Flood maps i) Chalk aquifer; ii) Greensand; iii) Limestone.

iii)

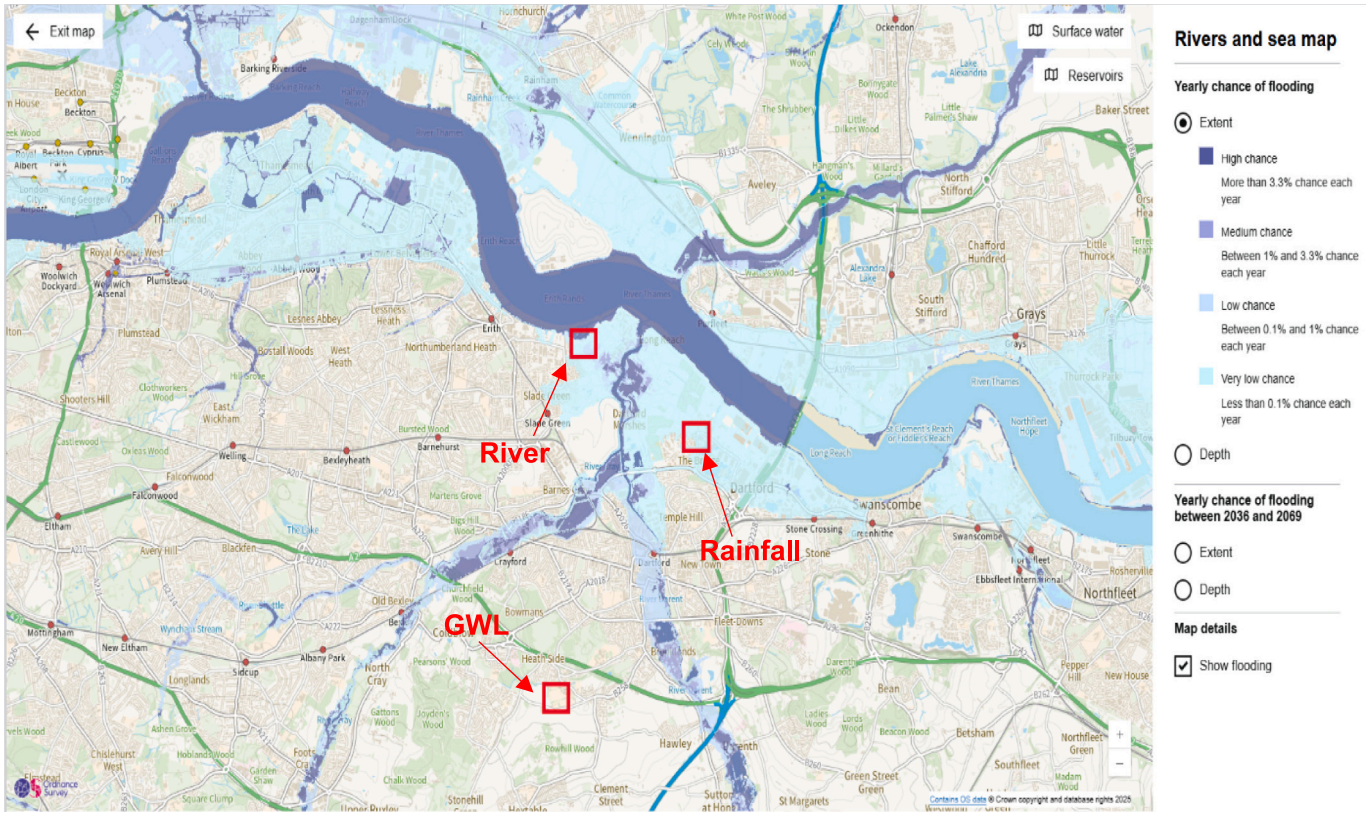


Fig. 2. (*continued*).

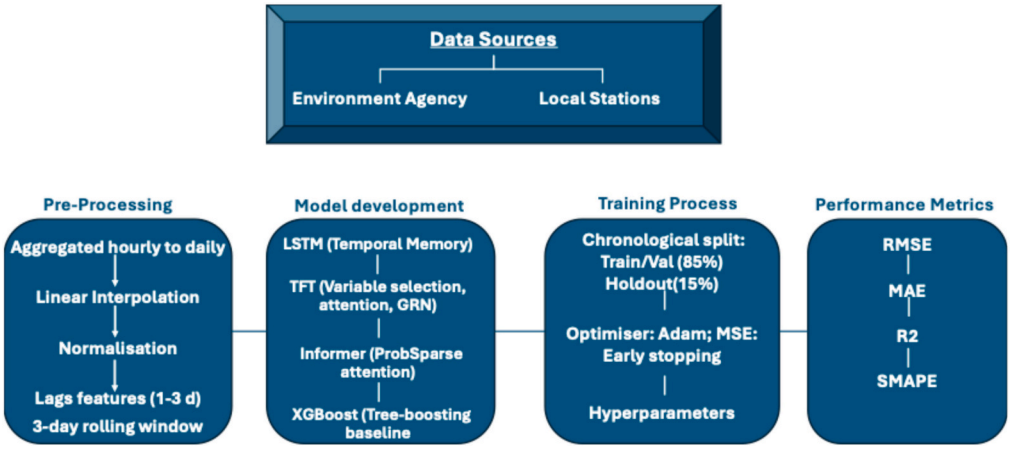


Fig. 3. A step-wise workflow of the study, showing data sources, preprocessing steps, training process, model development and performance evaluation.

long-sequence forecasting. It uses ProbSparse self-attention and a distillation operation to decrease computing costs (Zhou et al., 2021). Informer addresses the scalability constraints of traditional transformers, allowing for extended input windows without incurring excessive memory needs. This makes it appropriate for hydrological applications where long-term rainfall and groundwater dynamics can affect river levels. Informer has been used effectively in meteorology, renewable energy forecasting, and hydrologic modelling (Tepetidis et al., 2024). Supplementary Information (S3) provides technical specifics such as ProbSparse attention equations and architectural parameters.

2.2.4. *Extreme Gradient Boosting (XGBoost)*

Extreme Gradient Boosting (XGBoost) is an ensemble tree-based strategy that uses boosting with regularisation to increase prediction accuracy whilst avoiding overfitting (Chen and Guestrin, 2016). It has become the standard baseline for hydrological modelling due to its resilience, speed, and capacity to manage missing variables (Gaffoor et al., 2022). Whilst lacking the sequential modelling capabilities of neural networks, XGBoost excels on organised tabular data and serves as a baseline for assessing the additional value of deep learning approaches. In this work, XGBoost acts as a non-neural comparator, allowing us to determine if complicated sequence models significantly outperform ensemble baselines. Supplementary Information (S3)

provides a summary of hyperparameters and implementation settings.

2.3. Training, validation and testing (holdout method)

In this study, a strict holdout validation methodology, created specifically for evaluating predicting accuracy under practical operating situations, was used (Cerqueira et al., 2020). Using the holdout approach, the dataset is divided into discrete subsets for testing, validation and training. About 85 % of the dataset was used for model training and validation, with the remaining 15 % acting as a holdout test set. The dataset was divided chronologically into two major sections. For time series forecasting tasks, this chronological splitting technique is crucial because it keeps data from leaking, maintains the temporal integrity of forecasts, and ensures the model assessment takes performance on completely unknown future data into account (Weytjens and De Weerdt, 2021).

Data was further randomly separated inside the first training validation segment, with 85 % of the subset going towards model training and the remaining 15 % going towards validation. Model generalisability was improved by explicitly optimising model hyperparameters, using early stopping conditions, and avoiding overfitting using a validation subset.

During training, the holdout testing subset (final 15 %) was completely isolated and not given exposure to the models. An objective assessment of each model's prediction capacity was given via performance evaluation on the last subgroup, which is essential for determining practical applicability, especially when dealing with different hydrological circumstances and forecasting horizons (1–5 days).

Using the holdout method has a number of benefits over other strategies, such as k-fold cross-validation. In particular, non-stationary hydrological data, where temporal dynamics and sequential dependencies are crucial and better suited for a holdout validation strategy (Chandel and Ghosh, 2021). It produced more accurate and rationally meaningful performance estimates by preventing any data leakage through rigorous chronological separation. Thus, the holdout method increases confidence in the model's ability to generalise to new data, which makes it ideal for Thames Basin flood forecasting scenarios.

2.4. Performance metrics

For popular statistical measures – the Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), Coefficient of Determination (R^2), and symmetric Mean Absolute percentage error (SMAPE) – were chosen to thoroughly emulate the prediction performance of the models utilised in this research. Together, these measures offer distinct insight into the performance of the model, allowing for a thorough assessment from several angles (Chicco et al., 2021; Li et al., 2025).

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

The extent of prediction mistakes was assessed using the Root Mean Squared Error (RMSE), which assigns more weight to bigger disparities. This is especially important when it comes to flood prediction, since major mistakes in estimating peak river levels can greatly impact emergency response plans and readiness. The Mean Absolute Error (MAE), which averages the absolute discrepancies between projected and observed river levels, was also used as a supplementary indicator to provide an intuitive grasp of model accuracy. MAE is simpler to grasp in practical situations since it handles all mistakes equally, unlike RMSE. It can be represented as follows:

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

To calculate the percentage of observed variability that the models could account for, the coefficient of determination (R^2) was also included. For accurate flood forecasts, R^2 shows how well each model reflects the general patterns and temporal dynamics of river behaviour. It can be calculated as follows:

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

Additionally, prediction accuracy was assessed as a relative error measure using the Symmetric Mean Absolute Percentage Error (SMAPE). SMAPE has the benefit of being scale-independent, which allows for efficient comparisons of model performance across various river levels, rainfall intensities, and groundwater conditions. It also equally accounts for under- and over-predictions. In hydrological forecasting studies, where measurement scales might differ greatly between stations or hydrological conditions, this feature is advantageous.

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

Together, these criteria provide a thorough and impartial framework for evaluation, enabling in-depth analyses of each model's forecasting accuracy, robustness, and dependability over a range of hydrological circumstances and forecasting horizons.

3. Results

3.1. Model performance

Model performance was assessed using RMSE, MAE, R^2, and SMAPE metrics for three aquifer types (Chalk, Limestone, and Greensand) and various predicting horizons (1–4 days) (Tables 1, 2, and 3). At shorter timeframes (1–2 days), all models performed comparably well for the Chalk aquifer (Station 1), with only slight variations in RMSE and MAE values (~0.02–0.04). At a 1-day horizon, LSTM performed somewhat better than the other models (R^2 = 0.80, SMAPE = 5.74 %), with Informer (R^2 = 0.79, SMAPE = 6.05 %) and TFT (R^2 = 0.77, SMAPE = 6.44 %) following closely behind. The performance of XGBoost was somewhat worse (R^2 = 0.72). All models performed worse over longer horizons (3–4 days) (R^2 between 0.48 and 0.61), although TFT's performance remained more stable, as evidenced by lower SMAPE values than Informer and XGBoost.

All models showed remarkable forecast accuracy in the Limestone aquifer (Station 2), with extremely low RMSE (0.02–0.04) and MAE

Table 1
Station 1.

Aquifer Type	Horizon	Metrics	Models			
			TFT	*LSTM*	*Informer*	*XGBoost*
Chalk	1	RMSE	0.03	0.03	0.03	0.03
		MAE	0.02	0.02	0.02	0.02
		R^2	0.77	0.80	0.79	0.72
		SMAP	6.44	5.74	6.05	6.73
	2	RMSE	0.04	0.03	0.04	0.04
		MAE	0.03	0.02	0.03	0.03
		R^2	0.66	0.72	0.70	0.69
		SMAP	8.16	6.60	6.89	7.18
	3	RMSE	0.04	0.04	0.04	0.04
		MAE	0.03	0.03	0.04	0.03
		R^2	0.61	0.60	0.52	0.54
		SMAP	8.53	8.78	9.93	9.23
	4	RMSE	0.04	0.04	0.05	0.05
		MAE	0.03	0.03	0.04	0.03
		R^2	0.62	0.62	0.48	0.48
		SMAP	7.46	8.19	10.32	10.08

Table 2
Station 2.

Aquifer Type	Horizon	Metrics	Models			
			TFT	*LSTM*	*Informer*	*XGBoost*
Limestone	1	RMSE	0.02	0.02	0.02	0.02
		MAE	0.01	0.01	0.01	0.01
		R^2	0.99	0.99	0.99	0.99
		SMAP	1.64	1.82	1.53	1.56
	2	RMSE	0.03	0.03	0.03	0.03
		MAE	0.02	0.01	0.0.02	0.01
		R^2	0.97	0.98	0.98	0.98
		SMAP	2.83	2.38	2.57	2.05
	3	RMSE	0.03	0.03	0.04	0.03
		MAE	0.02	0.02	0.03	0.02
		R^2	0.97	0.97	0.95	0.97
		SMAP	2.74	2.80	5.33	2.72
	4	RMSE	0.04	0.04	0.05	0.04
		MAE	0.02	0.02	0.03	0.02
		R^2	0.95	0.96	0.94	0.95
		SMAP	3.54	3.38	4.24	3.33

Table 3
Station 3.

Aquifer Type	Horizon	Metrics	Models			
			TFT	*LSTM*	*Informer*	*XGBoost*
Greensand	1	RMSE	0.06	0.06	0.06	0.06
		MAE	0.06	0.04	0.04	0.04
		R^2	0.15	0.19	0.19	0.10
		SMAP	6.02	5.60	5.86	6.09
	2	RMSE	0.06	0.07	0.08	0.06
		MAE	0.04	0.04	0.06	0.04
		R^2	0.06	0.02	−0.33	0.11
		SMAP	6.55	6.63	8.57	6.06
	3	RMSE	0.07	0.07	0.04	0.07
		MAE	0.04	0.03	0.05	0.04
		R^2	0.00	0.05	−0.09	0.05
		SMAP	6.91	6.29	7.04	6.20
	4	RMSE	0.07	0.07	0.07	0.07
		MAE	0.05	0.04	0.04	0.04
		R^2	−0.17	0.00	−0.03	0.02
		SMAP	7.30	6.52	6.83	6.33

(0.01–0.02) values across all horizons. High R^2 values (≥0.94) demonstrated superior model-to-model prediction ability. With significantly lower SMAPE values (1.53–1.64 % at 1-day horizon), TFT and Informer performed marginally better than other models at shorter horizons, indicating a good model capability to reflect the distinct and responsive temporal hydrological processes typical of limestone aquifers.

In contrast to the Chalk and Limestone aquifers, all models showed poorer forecasting ability for the Greensand aquifer (Station 3). Particularly at horizons longer than two days, the RMSE and MAE were significantly higher (RMSE: 0.06–0.07, MAE: 0.04–0.07) with low or negative R^2 values. TFT showed negative R^2 values (−0.17) at the 4-day horizon, suggesting a significant decline in predictive power. The complexity and delayed hydrological response of the Greensand aquifer are demonstrated by the difficulties faced by Informer, LSTM, and XGBoost, which cast doubt on model predictions over longer time horizons (Shand et al., 2003b). All things considered, the performance study showed distinct model and aquifer-specific variations. The ability of transformer-based models (TFT and Informer) to capture temporal relationships is demonstrated by their higher performance, especially in sensitive aquifer settings like Limestone. The worst performance of all models in Greensand aquifers, however, points to the necessity of further modifying modelling techniques to better account for the slower and less responsive hydrological interactions that are characteristic of these geological environments. These comparison results show that, especially over longer time horizons, transformer-based models are consistently more resilient than ensemble techniques. When long-term dependencies were needed for delayed aquifer responses, XGBoost performed poorly, but LSTM demonstrated lasting utility as a sequential baseline. This demonstrates that rather than being presumed to be generally transferable, model design needs to be adapted to hydrological settings.

The comparison findings show distinct variations among aquifer types and between model designs. XGBoost was regularly outperformed by TFT and Informer, especially in areas with quick groundwater-river reactions. In the Limestone aquifer, for instance, TFT obtained R^2 values of 0.98 as opposed to 0.93 for XGBoost, highlighting the importance of attention processes in identifying transient hydrological correlations. Additionally, LSTM performed well, confirming its proven dependability for sequential hydrological data. However, XGBoost performed poorly in situations when long-term dependencies were necessary due to delayed aquifer responses, underscoring the limitations of tree-based methods in this regard.

3.2. Aquifer-specific performance

Model performance in Chalk, Limestone, and Greensand aquifers was compared, revealing significant variation that was mostly driven by the hydrogeological characteristics and groundwater-river interactions particular to each aquifer type. Even over longer forecasting horizons, the limestone aquifer (station 2) showed the highest overall model performance out of the three aquifer types, with RMSE values consistently as low as 0.02 and usually high R^2 values (~0.95–0.99) across all models (Table 2). Notably, this aquifer's historical data spanned the longest time of the three other stations, from April 2011 to January 2025. The prolonged time series most likely contributed significantly to model performance by providing adequate historical variability, such as several seasonal cycles and hydrologic extremes. These consistently correct predictions imply that limestone aquifers have quick and distinct groundwater-river interaction, which makes it possible for models, especially TFT and informer, to efficiently learn different temporal patterns (Oberhelman et al., 2024). Generally speaking, limestone aquifers exhibit very high permeability and quick groundwater recharge, which causes quick and noticeable changes in river levels after rainfall events (Neumann et al., 2003). This clarity in response makes accurate forecasts easier, due to hydrological processes that are predictable and simple to represent within the temporal framework of transformer-based models.

On the other hand, the greensand aquifer (station 3) showed noticeably worse predictive ability in all models, with considerably lower or negative R^2 values and higher RMSE and MAE values, especially at longer forecasting horizons (3–4 days). Despite a large historical dataset (April 2011 to April 2024) equivalent in duration to the Limestone aquifer, the Greensand aquifers have slower groundwater circulation and longer groundwater storage. Since this aquifer has a large storage capacity and delayed groundwater reaction, the poor performance suggests significant uncertainty and low prediction ability. These traits result in diffuse, delayed groundwater-river interactions, which makes it more difficult for models to predict changes in river level effectively, particularly when they depend mostly on short-term input features. These findings support a crucial finding: to improve predictive accuracy in hydrologically complex aquifers, standard temporal modelling techniques – even sophisticated ones like TFT and Informer – may need considerable modifications or extra data (like longer historical groundwater series or soil moisture data).

The chalk aquifer (station 1) performed at the middle levels. The historical data ranged from April 2011 to January 2024, offering a large dataset, slightly shorter than that of the Limestone aquifer. Models performed well for shorter predicting horizons (1–2 days) (R^2: 0.77–0.80). However, predicted accuracy fell considerably for time horizons longer than 2 days. Because of their high permeability, chalk aquifers often have reasonably fast groundwater-river interactions, although these interactions can be nuanced and complicated over time

(MacDonald and Allen, 2001; Ali et al., 2024). The intermediate dataset length, whilst substantial, may not completely reflect all complexities or long-term trends found in chalk aquifers, particularly under changing climatic conditions or major hydrological events, resulting in lower forecasting accuracy at longer time horizons.

The aquifer-specific research revealed important insights into flood predictions. First, the duration of historical datasets has a considerable impact on model accuracy, particularly for aquifers with distinct groundwater-river interactions (Limestones). Second, for aquifers with slow or diffuse groundwater reactions (Greensand), more historical data alone may not be enough; instead, specialised modelling techniques or hybrid approaches may be required. Finally, even aquifers with typically responsive groundwater dynamics, such as Chalk, may require longer or more extensive datasets to adequately capture complicated temporal patterns that influence long-term forecasting accuracy. This research significantly supports aquifer-specific forecasting methodologies, emphasising the importance of explicitly considering both hydrological qualities and dataset characteristics in future model development and flood risk management approaches.

These results complement the larger body of research on ML in flood forecasting, but they also go beyond it. Physical realism has been enhanced by recent developments, including surrogate modelling (Donnelly et al., 2022) and physics-informed neural networks (Donnelly et al., 2024). However, the majority of these studies concentrate on hydrological processes at the surface and hardly ever take aquifer-specific responses into account. Our findings show that predicting capability is highly influenced by the geological environment in addition to algorithmic design. Even the most sophisticated models can become useless if delayed subsurface processes are ignored, as evidenced by the extremely low performance seen in the Greensand aquifer ($R^2 \leq 0$).

3.3. Actual vs predicted

Fig. 4 supplies the visual comparison between the actual and predicted river levels for the chalk aquifer at forecasting horizons of 1 and 2 days using the Informer, LSTM, TFT and XGBoost models. These charts show the effects of longer prediction horizons and offer comprehensive insights into the forecasting capabilities and shortcomings of each model. At a 1-day horizon, numerical metrics in the ($R^2 = 0.79$ and 0.80, respectively) showed that both the Informer and LSTM models performed marginally better at the 1-day horizon, properly capturing river level oscillations and closely matching actual peaks and troughs. Minor but significant discrepancies in recording severe flow events were evident in the informer and LSTM models, which performed marginally better at the 1-day horizon, properly capturing river-level oscillations and closely matching actual peaks and troughs. Minor but significant discrepancies in recording severe flow events were evident in the informer's occasional modest overestimations during mild peaks and LSTM's slight underestimations of certain greater peaks.

The TFT model produced visually solid predictions that closely mirrored real river-level patterns with few variations, although quantitatively lagging behind LSTM and Informer ($R^2 = 0.77$). Although it is not numerically more accurate than LSTM or Informer, its performance at this horizon highlights its capacity to model short-term oscillations adequately. The XGBoost model, on the other hand, performed noticeably worse at the 1-day horizon, especially during peak-flow occasions, where it consistently and obviously underestimated. Despite the computational benefits, a significant drawback of simpler, tree-based approaches for short-term flood forecasting was their inability to capture short-term river-level peaks reliably.

At 2-day predicting horizons, all models showed decreased accuracy compared to the shorter horizons, indicating growing uncertainty. In terms of quantitative performance, the LSTM model performed best

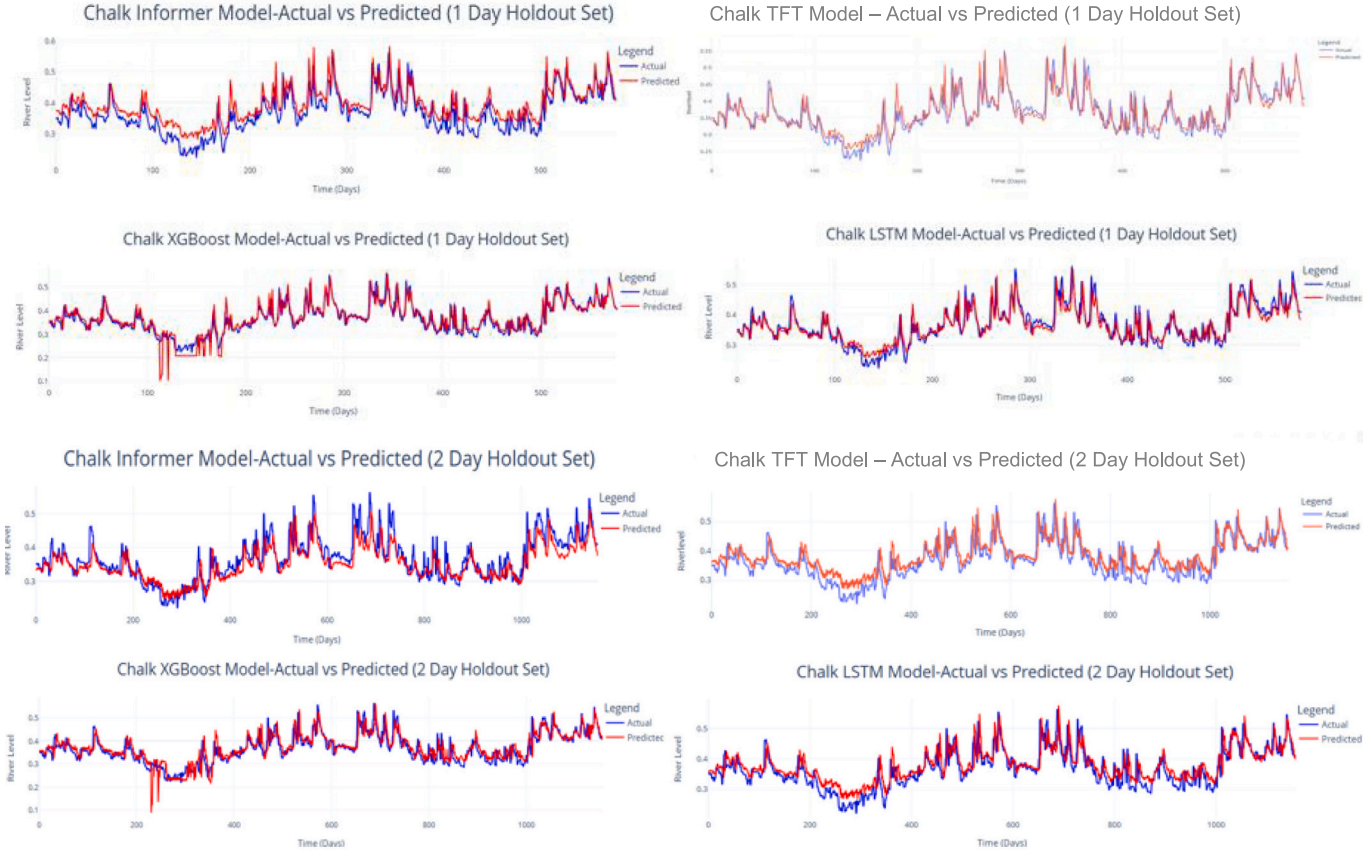


Fig. 4. Horizons 1 and 2 in the Chalk aquifer.

overall (R^2 = 0.72), closely followed by Informer (R^2 = 0.70). Both models successfully captured significant temporal patterns, but they revealed more deviations, especially during higher peak flows. At this longer horizon, the informer's propensity for modest overestimations becomes more noticeable.

In comparison to LSTM and Informer, the TFT model demonstrated somewhat lower quantitative accuracy at the 2-day horizon ($R^2 = 0.66$). Visually, TFT was still able to follow the overall temporal pattern successfully, but small deviations from actual values became more apparent, underscoring the difficulties that come with large predicting intervals. Compared to its 1-day performance, the XGBoost model showed somewhat better stability over the 2-day horizon; nonetheless, significant underestimations during peak events persisted. This demonstrates the continued inability of the more straightforward ensemble technique to capture intricate temporal dynamics that are essential to precise flood prediction.

When compared to more straightforward tree-based techniques like XGBoost, transformer-based models (Informer LSTM, and TFT) consistently showed greater predictive capacity. Although TFT maintained its visual reliability across horizons despite somewhat lower quantitative measures, LSTM and Informer demonstrated superior numerical accuracy at shorter horizons. The obvious decrease in prediction accuracy with longer forecast horizons highlights the need for models that can effectively manage uncertainty in flood forecasting situations.

The TFT model had the best overall accuracy at the 3-day horizon, with an R^2 value of 0.61. The LSTM model had a strong predictive performance (R^2 = 0.60). Both models correctly reproduced the fundamental river-level variations and temporal patterns, while minor differences, notably in reliably anticipating peak magnitudes and timing, grew when compared to shorter forecasting periods. The Informer model's accuracy decreased (R^2 = 0.52), resulting in overestimations during peak river-level occurrences. The XGBoost model struggled to accurately capture complicated hydrological processes over long time periods (R^2 = 0.54), frequently underestimating peaks (Fig. 5).

At the 4-day horizon, the TFT and LSTM models had the best prediction accuracy, with R^2 values of 0.62. While both models showed greater departures from real river levels than shorter periods, they nonetheless accurately reflected broad hydrological trends. Specifically, LSTM performed admirably, while variations at peak flow magnitudes were more noticeable. TFT produced similarly accurate overall forecasts, but with significantly higher fluctuation in peak and trough magnitudes than shorter horizons. Informer and XGBoost saw considerable performance declines at the 4-day horizon, with R^2 values of 0.48. Informer routinely overestimated river peaks, whilst XGBoost constantly showed instability and significant underestimations, highlighting their shortcomings in modelling long-term temporal dynamics.

Overall, examining model performance over horizons ranging from 1-day to 4-day predictions demonstrates a steady and progressive decrease in prediction accuracy as forecast intervals get longer. Transformer-based models (Informer and TFT) and the LSTM outperformed XGBoost in terms of sustaining performance levels over time. Nonetheless, the overall fall in accuracy highlights the complexities of long-term hydrological forecasting. Thus, for operational flood forecasting in Chalk aquifer contexts, TFT and LSTM models are particularly recommended due to their demonstrated robustness over longer time horizons, with potential accuracy gains achievable through improved modelling strategies or the incorporation of additional hydrological data.

Fig. 6 assesses and compares the performance of the Informer, LSTM, TFT and XGBoost models for the Greensand aquifer at 1- and 2-day forecasting horizons, using both quantitative measures and visual

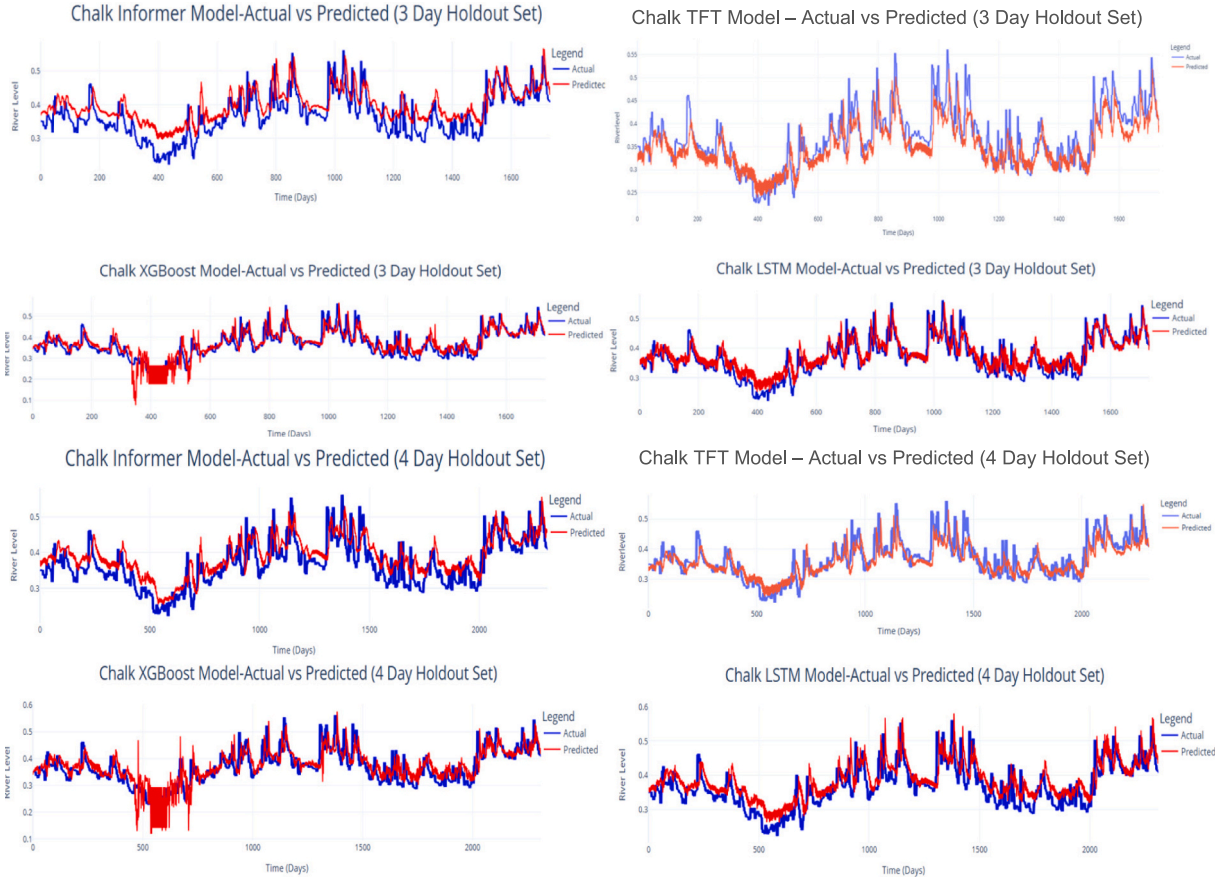


Fig. 5. Horizons 3 and 4 in the Chalk aquifer.

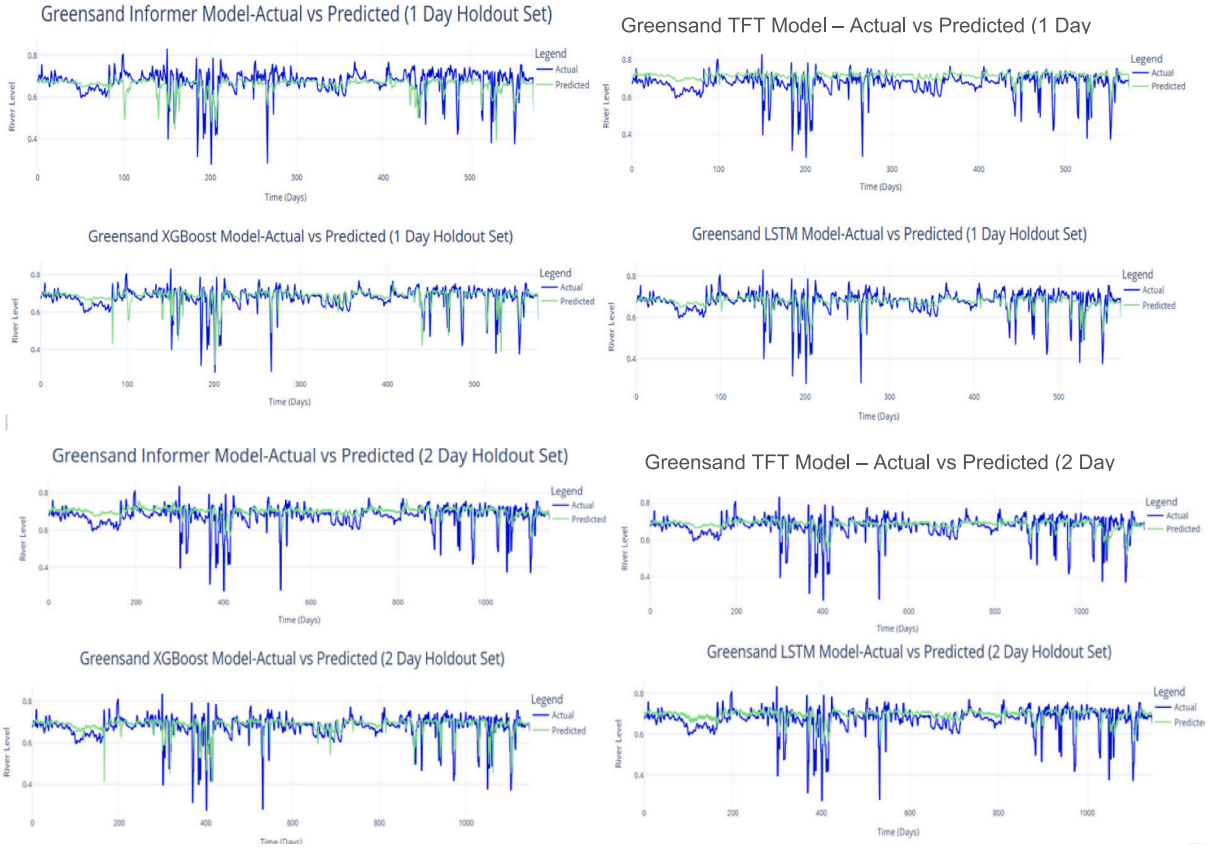


Fig. 6. Horizons 1 and 2 in the Greensand aquifer.

forecasts. All models had low R^2 values at the 1-day horizons as shown in Table 3, ranging from 0.10 (XGBoost) to 0.19 (LSTM and Informer), indicating inadequate forecasting accuracy. This is seen graphically by the notable variations in peak and trough prediction across all models in Fig. 6. Even though there were significant errors in peak magnitudes, the LSTM and Informer models outperformed TFT and XGBoost in terms of timing some river-level fluctuations. Significant limits were shown by the TFT model ($R^2 = 0.15$), which consistently underestimated peaks and was unable to describe abrupt changes adequately. The least accurate model ($R^2 = 10$), XGBoost, continuously underestimated peak river levels, demonstrating its incapacity to manage the more intricate groundwater-river interactions seen in the Greensand aquifer. Due to the complicated and slow groundwater reaction that results in less predictable surface water interactions, aquifers such as greensand are challenging to simulate for river dynamics (Shand et al., 2003b), as seen by their comparatively poor performance measures.

For the majority of models, predicted accuracy further decreased at the 2-day forecasting horizons. Despite maintaining a very low overall accuracy, the LSTM remained somewhat steady and achieved the greatest R^2 value (0.06). In terms of catching peak flows, LSTM forecasts show consistent but marginally higher errors as seen in Fig. 6. The TFT model struggled with peak and trough prediction, which maintained poor accuracy ($R^2 = 0.06$). Due to significant peak misalignment, severe prediction errors, and consistent overestimation or underestimation, the Informer's accuracy significantly decreases (R^2 = -0.33). Significant challenges in modelling temporal dynamics at this horizon are highlighted by the negative R^2 values, which show performance below a simple average forecast. With its continued severe peak underestimation and minimal predictive improvement, XGBoost continued to perform quite poorly ($R^2 = 0.06$).

Overall, compared to the Chalk aquifer, the Greensand aquifer results at the 1-day and 2-day horizons clearly show significant modelling challenges. This is primarily because of the unique hydrological characteristics of greensand, which include complex surface-subsurface interactions and slower groundwater responses. Peak forecasts were significantly difficult for all models, indicating that more complex modelling techniques or more data sources are needed to capture groundwater-river interactions in detail. Out of all the models that were examined, the LSTM had somewhat higher, but restricted, accuracy across these short horizons. Informer came in second at the 1-day horizon. Due to inherent aquifer-specific complexity, transformer-based models did not significantly outperform simpler models in this case, despite their potential advantage in capturing complicated temporal dynamics. Future model improvements could be on using models that are especially tailored to the particular groundwater dynamics of greensand-type aquifer or on integrating more comprehensive hydrological data.

Fig. 7 shows the capacity of the models to estimate Greensand aquifer river levels during extended prediction intervals of three and four days. Due to its large storage capacity and delayed river-groundwater interactions, which become more noticeable over longer horizons, the Greensand system poses unique hydrological issues. Although it was still low, the LSTM model had the greatest R^2 value (0.05) at the 3-day horizon, suggesting only limited explanatory ability. However, with a negative R^2 (−0.09), the Informer model did not perform well, indicating that it was unable to acquire meaningful temporal characteristics under the growing uncertainty that comes with longer forecasts in this aquifer. TFT likewise had trouble, providing no discernible improvement over the basic baseline ($R^2 = 0.00$).

All of the models were unable to precisely match the real river-level peaks and troughs, which were evident in Fig. 7. With an R^2 of 0.05, XGBoost equalled LSTM despite having a simpler structure and exhibited somewhat more consistent trend-following behaviour. SMAPE values (<6) were elevated in all models, nevertheless, suggesting that

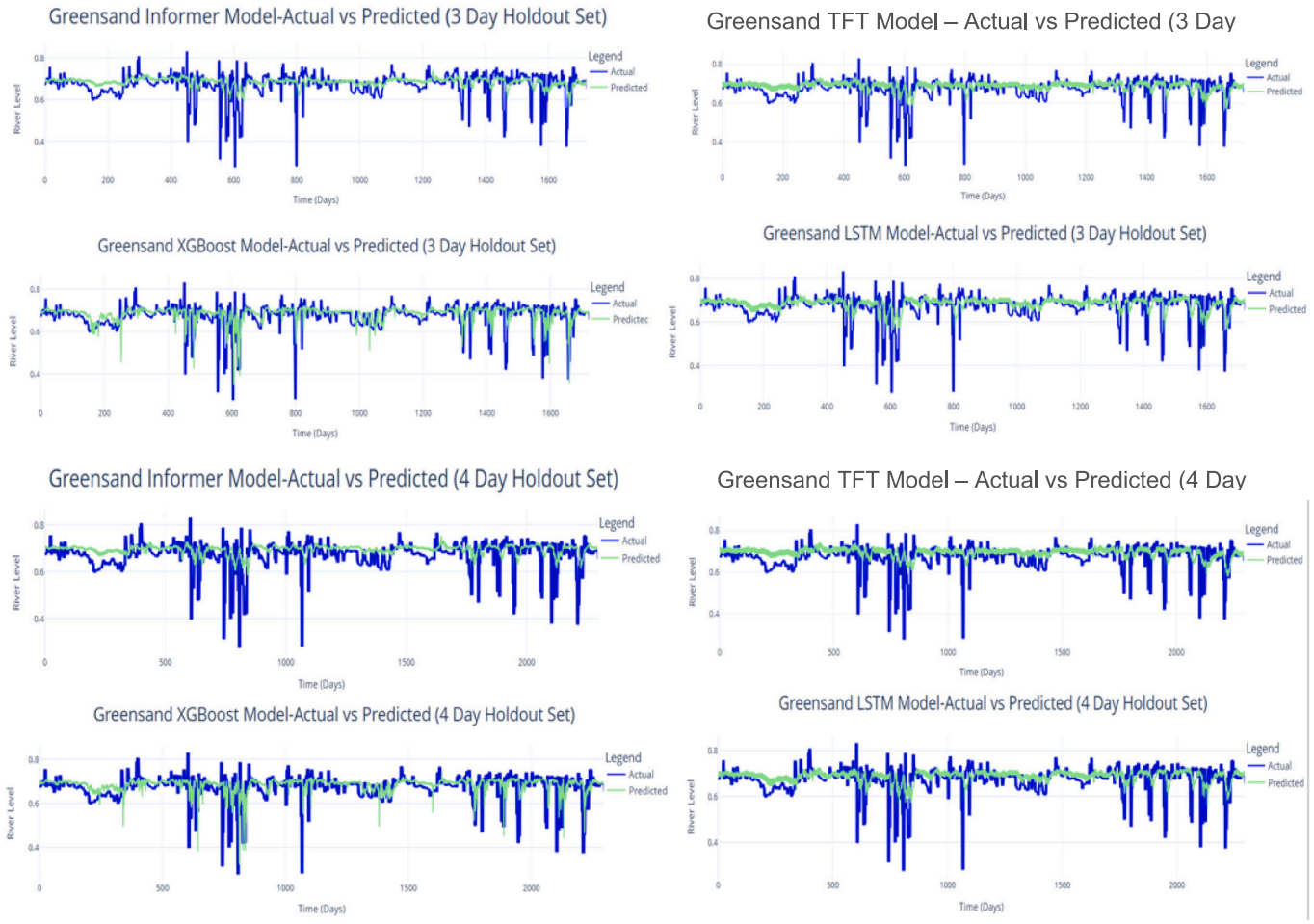


Fig. 7. Horizons 2 and 3 in the Greensand aquifer.

relative error increased with flow conditions. Interpretability was negatively impacted by frequent noise and peak misplacement, which were specifically introduced by Informer.

By a 4-day horizon, all model performance further declined, with a very negative R^2 value of −0.17; the TFT model proved unable to capture even the most fundamental river-level patterns in Greensand over a long period of time. During critical peak moments, the Informer model (R^2 = -0.03) displayed chaotic predictions that deviated from the actual flow profiles, continuing its dismal performance. In accordance with its continual reduction over time, LSTM's R^2 fell to 0.00. Its utility was restricted at this range by its inability to mimic fluctuations and extremes, even if its visual output still roughly mirrored the overall direction of river level as seen in Fig. 7. XGBoost's usefulness for practical flood forecasting was diminished since it continued to underestimate peak and over-smooth the signal, whilst being comparatively more stable (R^2 = 0.07). The SMAPE increased to 6.3–7.3 for all models, suggesting significant distortion in the predictions. This was verified visually, since there were significant discrepancies between the lines that were seen and those that were projected during times of hydrological activity.

The Greensand aquifer's 3-day and 4-day forecasting horizons demonstrated that this setting offers the most predictive difficulty among the three aquifer types. Regardless of architecture, the models did not generalise successfully. Greensand's groundwater movement is slow and diffuse, which probably reduces the accuracy of surface-level indicators like rainfall or river levels (Shand et al., 2003b). In contrast to transformer-based models (TFT and Informer), which seem to need deeper temporal signals than greensand systems can provide within the observed characteristics, LSTM and XGBoost were comparatively more consistent and stable, despite their overall poor performance. Specifically, Informer was extremely sensitive to horizon extension, displaying unpredictable or overfitting results after one day. To enhance performance in slower aquifer systems like greensand, our results emphasise the need for aquifer-specific model selection and maybe the incorporation of extra variables (such as soil moisture, baseflow indices, or groundwater recession rate).

In the Limestone aquifer, all models demonstrated consistently high prediction performance, especially at the 1-day and 2-day timeframes. Because of its sensitive and well-drained characteristics, which provide a more distinct link between hydrological inputs (such as rainfall and groundwater levels) and river level output, this aquifer type demonstrated the best accuracy across all stations.

All models performed almost perfectly at the 1-day horizon with R^2 values of 0.099, as shown in Table 2 for each model. This suggests a system with a limited horizon that is quite predictable. Compared to other models, the Informer model was the most accurate, as seen by its lowest SMAPE (1.53). With SMAPE ratings of 1.64 and 1.56, respectively, TFT and XGBoost were in close pursuit, whilst LSTM lagged somewhat (SMAPE = 1.82), but it was still achieving very good results. All models closely followed the real peak and trough of the river level, with very little variance, based on the visual charts. Excellent fit and little lag were shown by the Informer and XGBoost models' very seamless alignment with the observed data. TFT likewise matched the river dynamics nicely, although at strong peaks, it responded a little more rounded, perhaps because of its multi-head attention smoothing function. LSTM demonstrated slight delays in capturing certain peak

occurrences, despite their perceived similarity.

Even though there were minor accuracy declines, all models continue to perform well at the 2-day forecast horizon, as shown in Fig. 8. Despite the increased uncertainty of making a prediction two days in advance, R^2 values stayed high, ranging from 0.97 to 0.98, indicating robust model generalisation. With the lowest SMAPE of 2.38 and an R^2 of 0.98, LSTM dominated the field and demonstrated exceptional consistency in monitoring river-level variations. Additionally, XGBoost demonstrated its ability to capture structured patterns in high-quality, well-behaved datasets such as this one by performing well ($R^2 = 0.98$, SMAPE = 2.05).

When compared to its 1-day horizon output, Informer showed a slightly higher propensity to overpredict high peaks, but it still produced solid predictions with $R^2 = 0.89$ and SMAPE = 2.57. TFT reported a high SMAPE (2.83), indicating a somewhat increased sensitivity to shifting dynamics over longer periods, even if it was still operating well ($R^2 = 0.97$). These measurements are supported by the plots, which continue to demonstrate extremely tight alignment between LSTM and XGBoost and real values, particularly around increasing limbs of hydrographs. Although TFT and Informer maintained solid timing, they may have overestimated some low-flow sectors because of their deeper temporal designs' reliance on lag patterns.

At 1- and 2-day timeframes, the limestone aquifer demonstrated remarkable predictability with all four models operating at or close to ideal levels. Even basic models like XGBoost performed well because of the close correlation between input characteristics and river levels, which was probably caused by the aquifer's moderate permeability and steady hydrological behaviour (Thakkar and Lohiya, 2022).

Fig. 9 assesses the performance of the Informer, TFT, LSTM, and XGBoost models for the Limestone aquifer across extended forecasting horizons of 3 and 4 days. This aquifer provides ideal circumstances for precise long-term forecasting due to its quick and direct groundwater-surface water interactions (Oubagaranadin et al., 2007). All four models showed consistently good predicted accuracy at the 3-day horizon. With an R^2 of 0.97, both the TFT and LSTM models were in the lead and had a very high capacity for explanation. Especially when it comes to capturing critical hydrological peaks, their low SMAPE values (2.74 and 2.80, respectively) demonstrate precise alignment with actual river-level changes. With very slight variations at peak flows, the figures demonstrate how TFT and LSTM closely reflect the observed river patterns.

Equally well, XGBoost generated predictions that closely matched the observed values ($R^2 = 0.97$). Although it occasionally underestimated peak magnitudes, its more straightforward structure was adequate in the Limestone environment and successfully captured peak dynamics. At this horizon, the informer model's accuracy was somewhat lower ($R^2 = 0.95$), and its SMAPE significantly increased to 5.33, suggesting more relative mistakes. Despite having adequate timeless for detecting river-level changes, informer visually exaggerated peak magnitudes, especially around significant hydrological events, indicating sensitivity to prolonged horizons.

Overall predictive quality remained quite good, although model accuracy started to exhibit mild reductions at the 4-day horizon. TFT, LSTM, and XGBoost all exhibited robust performance ($R^2 = 0.96$), suggesting reliable forecasts despite longer lead periods. Even across extended forecasting horizons, LSTM consistently shows its capacity to capture complex temporal dynamics by achieving the lowest SMAPE (3,38). Even whilst TFT was still quite accurate (SMAPE = 3.54), it showed more flattering peaks and troughs, which indicated that it was challenging to capture abrupt changes over a long period of time adequately.

In constant hydrological circumstances, such as those seen in

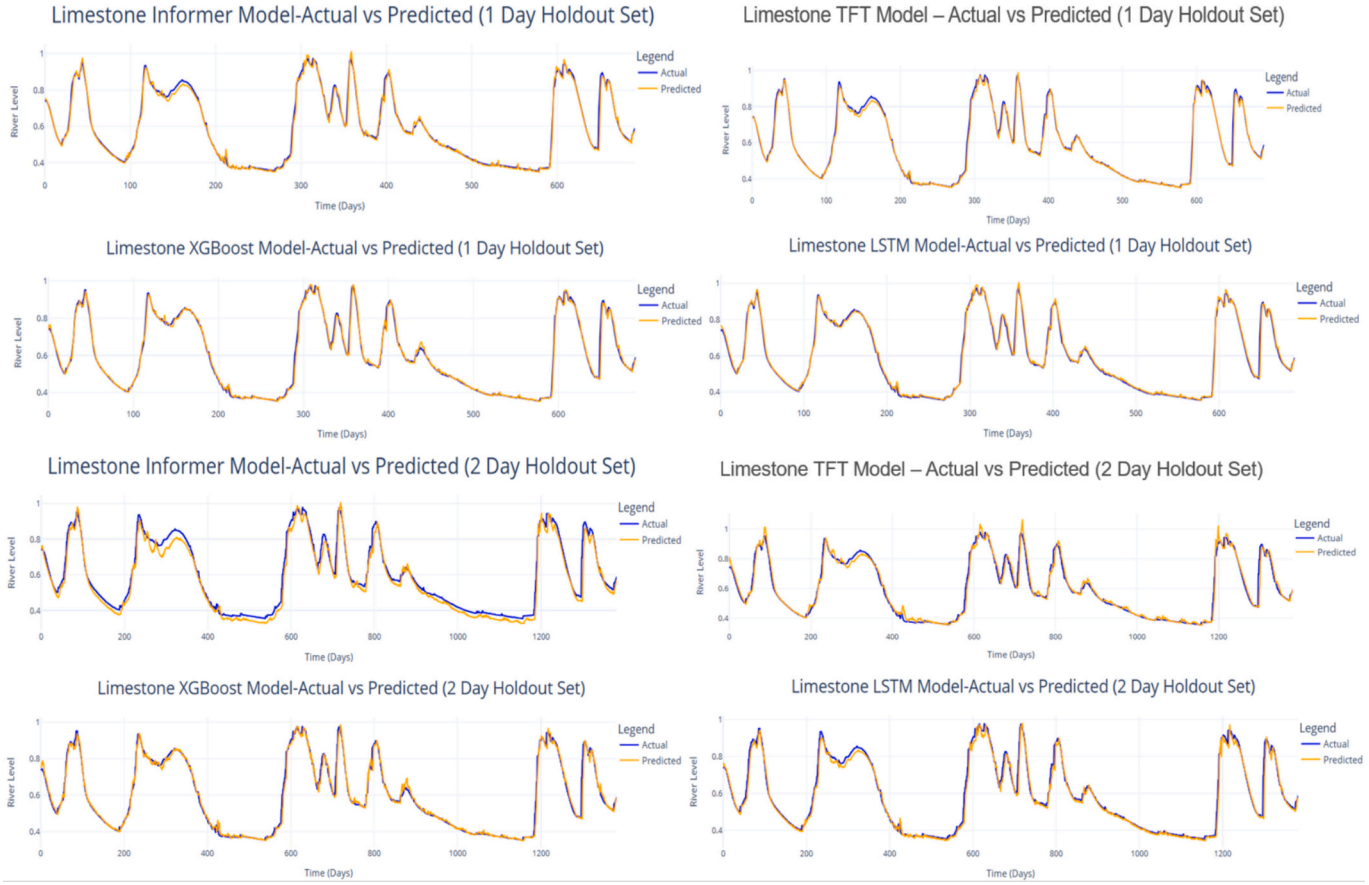


Fig. 8. Horizons 1 and 2 in the Limestone aquifer.

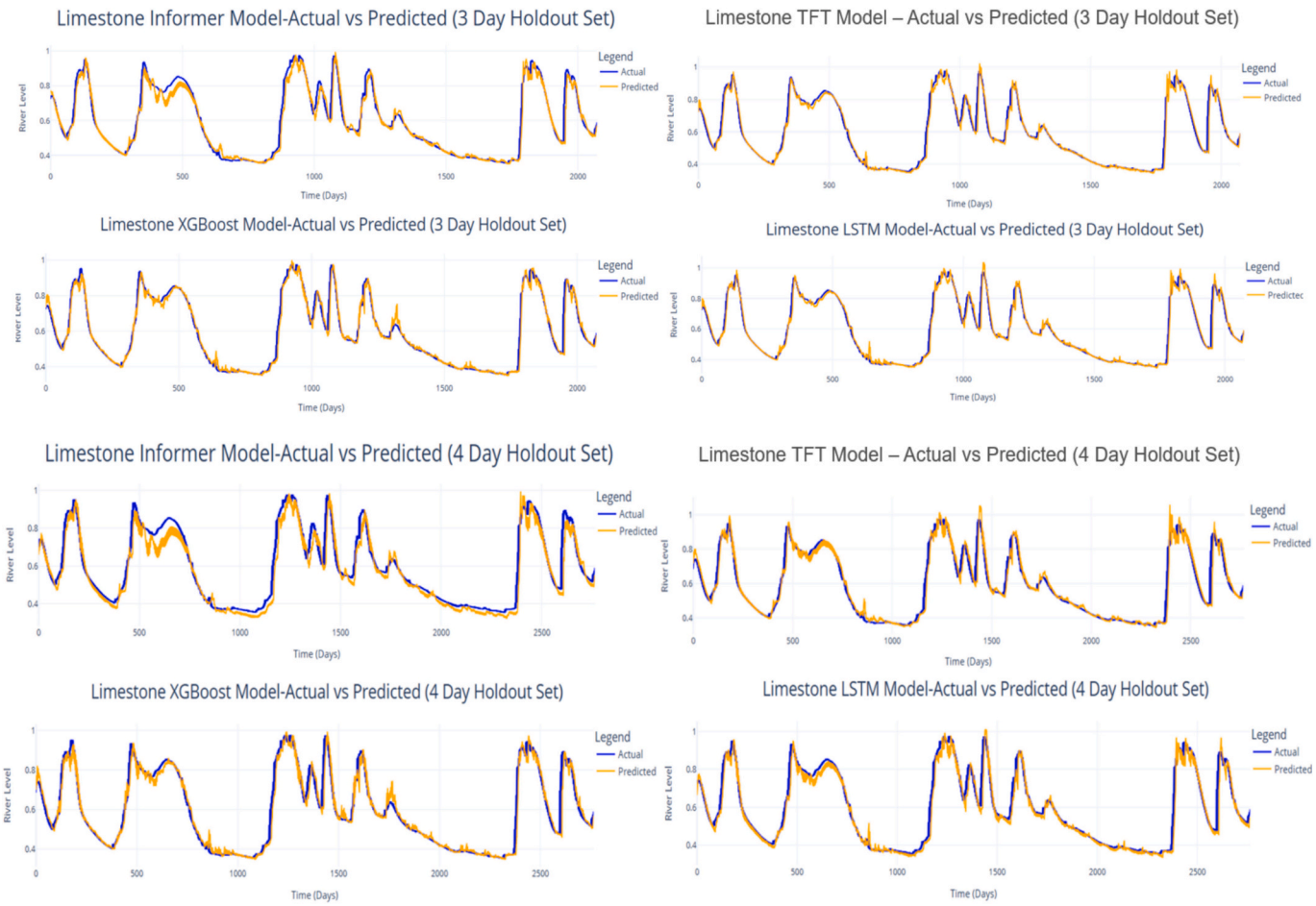


Fig. 9. Horizons 3 and 4 in the Limestone aquifer.

limestone aquifers, XGBoost demonstrated exceptional efficacy by maintaining competitive performance (SMAPE 3.33). Its forecasts showed little latency and accurately captured notable peaks and troughs, nearly matching real river levels. On the other hand, the Informer model showed a more significant decline in performance (R^2 = 94, SMAPE = 4.24 %), which was especially noticeable in over-estimations during high flow and significant departures during low flow. Informer's decreased accuracy shows the model's susceptibility to cumulative prediction mistakes over this extended horizon.

Overall, because of its hydrological features that encourage quick groundwater river interactions, the limestone aquifer enabled outstanding model performance even at longer horizons. The TFT and LSTM models performed better in managing more prediction uncertainty over longer time horizons. In highly predictable aquifer systems, complicated models may not necessarily deliver considerable gains, as demonstrated by XGBoost, which offered a simpler but equally reliable option. Informer showed decreased stability as forecasting intervals grew, although it was robust at shorter horizons. This suggests that it should not be used for longer-term operational forecasting unless it is better calibrated or backed by additional data sources.

This study's integration of aquifer-specific characteristics into a comparative ML framework represents a significant methodological advancement. Our method specifically takes into account groundwater-river interactions across several aquifers, in contrast to previous flood forecasting studies that frequently consider catchments as homogenous entities. This is the first proof that ML performance in flood forecasting is consistently aquifer-dependent that we are aware of. Therefore, this methodology goes beyond algorithmic benchmarking to show how forecast reliability is governed by geological variability.

3.4. Aquifer-specific interactions

The heat map in Fig. 10 outlines the essential requirement for aquifer-specific approaches to flood prediction modelling, which shows different groundwater-river interactions and rainfall effects within the three aquifer types: limestone, chalk, and greensand.

The Limestone aquifer showed a strong positive correlation (r = 0.84) between groundwater and river levels, due to its karstic nature, which is marked by high permeability and quick groundwater transfer (Ali et al., 2024). These circumstances allow for rapid groundwater recharge and release, which has a major effect on river levels nearly immediately after groundwater variations. Rainfall showed a poor association (r = 0.067), suggesting that groundwater dynamics, rather than direct rainfall-runoff reactions, are the primary cause of river level fluctuation. This explains why models, particularly transformer-based models (TFT and Informer) and LSTM, which effectively take advantage of the strong groundwater-river link to predict river dynamics, consistently have higher prediction accuracy.

The Chalk aquifer has a moderate correlation (r = 0.26) between groundwater and river levels. Compared to limestone, this mild association suggests that groundwater contributions to river flow are less immediate or strong, but still significant. Rainfall showed an even less correlation (r = 0.12) with river levels, indicating that other hydrological processes, such as storage capacity and delayed reactions within the aquifer, may complicate groundwater dynamics, even though it is relatively influential. Over the course of the forecasting horizons, these dynamics led to moderately varied model results. Due to their ability to capture subtle interactions over extended time periods, transformer-based models and LSTM continued to perform better than more

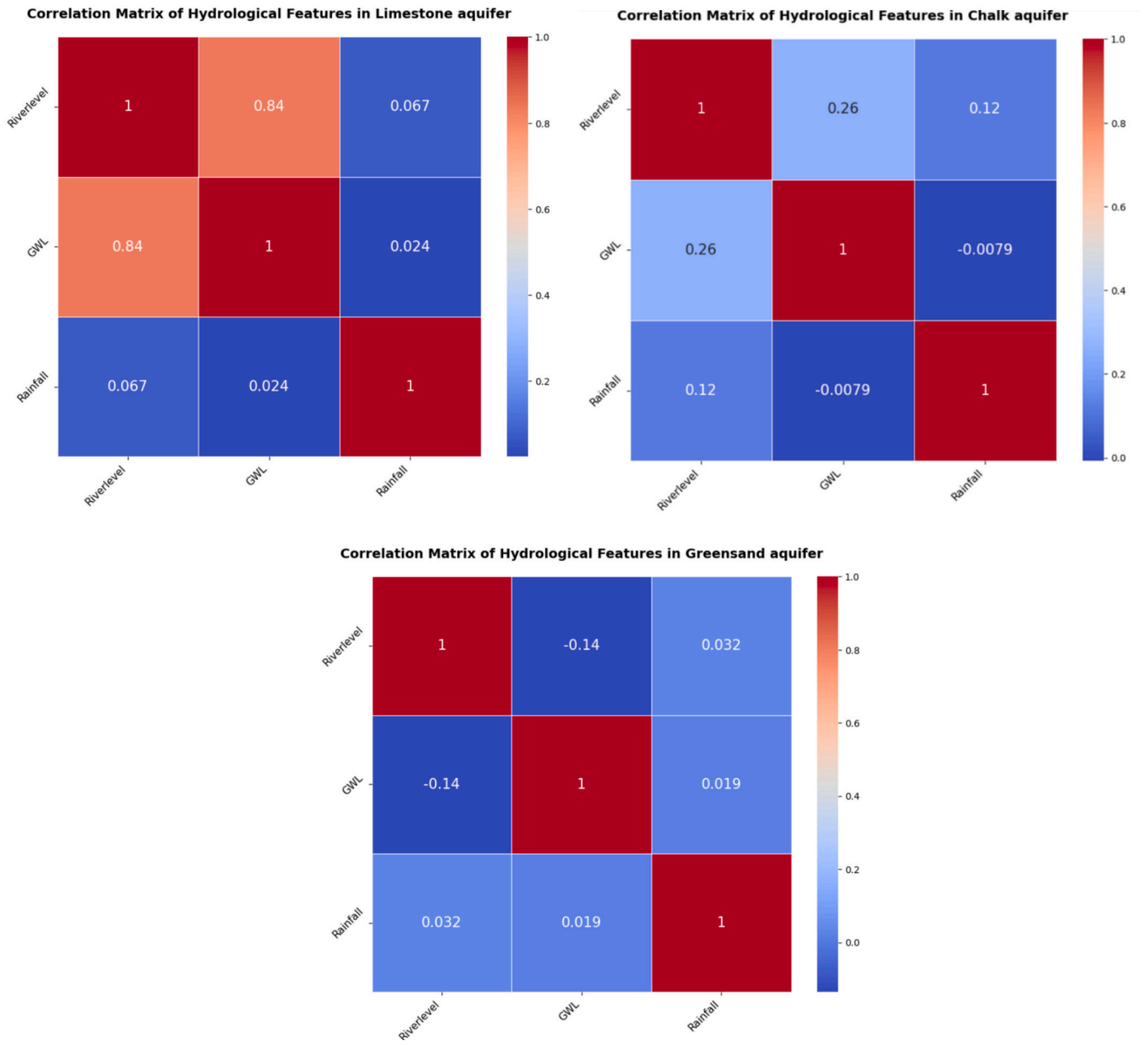


Fig. 10. Heat map of all aquifers.

straightforward models like XGBoost.

Moreover, in the Limestone aquifer, a weak negative association ($r = -0.14$) between river levels and groundwater in the Greensand aquifer. This negative association points to more intricate hydrological processes that may involve substantial groundwater storage and delayed release, which might mitigate or even reverse immediate river-level changes. Rainfall and river levels showed a weak connection ($r = 0.032$), highlighting the rainfall's limited direct impact and making flood forecasting in the Greensand aquifer even more challenging. Significant difficulties in effectively modelling flood dynamics are highlighted by the intricacy and indirect interactions in Greensand environments, which resulted in noticeably reduced forecast accuracy across all models and horizons.

These results highlight the need to modify flood prediction techniques to account for particular aquifer properties. Since groundwater-river interactions occur quickly in limestone aquifers, developed models that appropriately take advantage of these robust linkages are obviously beneficial. Models incorporating extra hydrological variables or extensive historical data may benefit chalk aquifers, which need careful consideration of somewhat delayed groundwater interactions. On the other hand, Greensand aquifers present considerable forecasting difficulties, which may necessitate integrated modelling techniques that account for intricate, delayed groundwater interactions. Enhancing flood risk management frameworks' prediction accuracy and dependability requires this aquifer-specific knowledge.

3.5. Implications for flood risk management

The study's conclusions have significant applications in improving flood early warning systems, especially when using cutting-edge machine learning algorithms. The significant variation between the three stations highlights the vital need for aquifer-specific modelling techniques to forecast flood occurrences precisely. Specifically, transformer-based models (TFT and Informer) and LSTM revealed higher prediction ability, successfully capturing complicated groundwater-river interactions and temporal hydrological patterns. Particularly in regions with quick groundwater-river exchange, like limestone aquifers, the

integration of these sophisticated models into current flood forecasting frameworks could greatly enhance early warnings, offering increased dependability and earlier lead times.

Moreover, the demonstrated capability of these sophisticated models to sustain accurate forecasts across many forecasting horizons makes them particularly ideal for practical usage, enabling flood control authorities to take pre-emptive actions several days in advance. This capacity is extremely useful since it enables quicker evacuation planning, focused resource allocation, and more educated reactions during flood disasters. Traditional or simpler models, such as XGBoost, whilst computationally efficient, demonstrated limitations in dealing with complex hydrological interactions, implying that advanced deep learning models are a better choice for operational flood forecasting.

Furthermore, by continually updating projections as new data becomes available, incorporating these predictive models into already-existing flood management systems might improve their responsiveness. Real-time forecasting using transformer-based models and LSTM might considerably increase preparation, decrease flood-related damage, and enhance public safety, particularly in high-risk zones characterised by aquifer-specific features. This work emphasises not only the predictive power of these sophisticated model but also their practicality and relevance in real-world flood risk management scenarios, opening the way for more robust and adaptable flood control systems.

Planning for resilience and flood control is directly impacted by the findings. With models achieving very high accuracy ($R^2 > 0.98$) in limestone aquifers, predictions can provide accurate short-range flood warnings. Chalk aquifers are more appropriate for preparation planning with thorough uncertainty communication because of their moderate predictability ($R^2 = 0.77–0.80$). In contrast, the Greensand aquifer's weak performance indicates the necessity for adaptive measures, such as hybrid modelling or improved monitoring, to prevent an over-reliance on ML predictions. These unique perspectives suggest the creation of a flood resilience framework that takes into account aquifers and matches forecasting capabilities with regional hydrological reality. This study offers an aquifer-aware framework that can be immediately used by organisations like the Environment Agency for customised flood management by clearly connecting the geological setting to model dependability.

4. Summary and conclusions

Using cutting-edge machine learning methods, to the best of the authors' knowledge, this work offers the first aquifer-specific, multi-horizon comparative investigation of flood forecasting in the Thames Basin. Through the explicit incorporation of groundwater-river interactions into the forecasting framework, we show that aquifer type has a significant impact on prediction. Dependability. This methodological contribution emphasises the significance of geological context in machine learning flood prediction, going beyond algorithmic benchmarking.

The findings demonstrated that transformer-based models (TFT and Informer) consistently performed better than XGBoost and LSTM, with intermediate skill in chalk ($R^2 \approx 0.77–0.80$) and perfect accuracy in limestone aquifers ($R^2 > 0.98$). The efficacy of existing machine learning techniques is limited by delayed subsurface reactions, as seen by the poor performance in Greensand aquifers ($R^2 < 0$). The novel insight that geological variability is a key factor in prediction robustness is added by these findings, which also support recent developments in physics-informed and hybrid models. This study emphasises the value of aquifer-specific modelling techniques, reaffirming that broad modelling approaches cannot handle the complexity of hydrological forecasting in various geological contexts. Future studies should improve these models even further, including more hydrological variables, and investigate hybrid or ensemble modelling techniques to deal with the inherent uncertainties found, especially in intricate environments like Greensand aquifers.

The study's limitations include its sole use of continuous forecasting without hybrid or physics-informed limits, its dependence on daily aggregated data, and its single-region emphasis. Future studies should investigate hybrid data-process frameworks, higher-resolution monitoring networks, and physics-informed machine learning techniques. They should also test the framework in a variety of hydrogeological environments. Notwithstanding these drawbacks, the architecture described here is obviously applicable to various aquifer-controlled basins around the globe. Water managers and organisations like the Environment Agency may create more dependable flood resilience plans that consider aquifers by customising flood forecasting techniques to the geological environment.

CRediT authorship contribution statement

Ali J. Ali: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Conceptualization. **Ashraf A. Ahmed:** Writing – review & editing, Supervision, Project administration, Investigation, Funding acquisition.

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 want to thank the editors and anonymous reviewers, whose comments greatly improved this manuscript. This study was partially funded by the UKRI project 10063665.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2025.180756.

Data availability

Data will be made available on request.

References

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.
Ali, A.J., Ahmed, A.A., 2024. Long-term AI prediction of ammonium levels in rivers using transformer and ensemble models. Clean. Water 2, 100051.
Ali, A.J., Ahmed, A.A., Abbod, M.F., 2024. Groundwater level predictions in the Thames Basin, London over extended horizons using transformers and advanced machine learning models. J. Clean. Prod. 484, 144300.
Bearcock, J.M., Smedley, P.L., 2010. Baseline Groundwater Chemistry: The Palaeogene of the Thames Basin.
Brunner, M.I., Slater, L., Tallaksen, L.M., Clark, M., 2021. Challenges in modeling and predicting floods and droughts: a review. WIREs Water 8 (3), e1520.
Cerqueira, V., Torgo, L., Mozetič, I., 2020. Evaluating time series forecasting models: An empirical study on performance estimation methods. Machine Learning 109, 1997–2028.
Chandel, V.S., Ghosh, S., 2021. Components of Himalayan river flows in a changing climate. Water Resour. Res. 57 (2) p.e2020WR027589.
Chen, T., 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 Computer Science 7, e623.
Crooks, S.M., Kay, A.L., 2015. Simulation of river flow in the Thames over 120 years: evidence of change in rainfall-runoff response? J. Hydrol. Reg. Stud. 4, 172–195.
Deepa, B., Ramesh, K., 2022. Epileptic seizure detection using deep learning through min max scaler normalization. Int. J. Health Sci. 6, 10981–10996.
Donnelly, J., Abolfathi, S., Pearson, J., Chatrabgoun, O., Daneshkhah, A., 2022. Gaussian process emulation of spatio-temporal outputs of a 2D inland flood model. Water Res. 225, 119100.

Donnelly, J., Daneshkhah, A., Abolfathi, S., 2024. Physics-informed neural networks as surrogate models of hydrodynamic simulators. Sci. Total Environ. 912, 168814.
Environment Agency, 2022. Thames River Basin District Flood Risk Management Plan 2021 to 2027. http://www.gov.uk/government/publications.
Fan, Y., 2024. Ensemble flood predictions for river Thames under climate change. Nat. Sci. Open 3 (1), 20230027.
Gaffoor, Z., Pietersen, K., Jovanovic, N., Bagula, A., Kanyerere, T., Ajayi, O., Wanangwa, G., 2022. A comparison of ensemble and deep learning algorithms to model groundwater levels in a data-scarce aquifer of Southern Africa. Hydrology 9 (7), 125.
Hamel, P., Tan, L., 2022. Blue–green infrastructure for flood and water quality management in Southeast Asia: evidence and knowledge gaps. Environ. Manag. 69 (4), 699–718.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9 (8), 1735–1780.
Hussain, F., Wu, R.S., Wang, J.X., 2021. Comparative study of very short-term flood forecasting using physics-based numerical model and data-driven prediction model. Nat. Hazards 107 (1), 249–284.
Jenkins, K., Hall, J., Glenis, V., Kilsby, C., 2018. A probabilistic analysis of surface water flood risk in London. Risk Anal. 38 (6), 1169–1182.
Kang, M., Tian, J., 2018. Machine Learning: Data Pre-processing. In: Prognostics and Health Management of Electronics: Fundamentals, Machine Learning, and the Internet of Things, pp. 111–130.
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.
Koya, S.R., Roy, T., 2024. Temporal Fusion Transformers for streamflow Prediction: Value of combining attention with recurrence. Journal of Hydrology 637, 131301.
Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A.K., Hochreiter, S., Nearing, G.S., 2019. Toward improved predictions in ungauged basins: exploiting the power of machine learning. Water Resour. Res. 55 (12), 11344–11354.
Le, X.H., Ho, H.V., Lee, G., Jung, S., 2019. Application of long short-term memory (LSTM) neural network for flood forecasting. Water 11 (7), 1387.
Li, D., Hu, J., Li, M., Zhao, S., 2025. A long-term dissolved oxygen prediction model in aquaculture using transformer with a dynamic adaptive mechanism. Expert Systems with Applications 259, 125258.
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.
MacDonald, A.M., Allen, D.J., 2001. Aquifer properties of the Chalk of England. Q. J. Eng. Geol. Hydrogeol. 34 (4), 371–384.
Motta, M., de Castro Neto, M., Sarmento, P., 2021. A mixed approach for urban flood prediction using machine learning and GIS. Int. J. Disaster Risk Reduction 56, 102154.
Neal, C., Neal, M., Hill, L., Wickham, H., 2006. The water quality of the River Thame in the Thames Basin of south/south-eastern England. Sci. Total Environ. 360 (1–3), 254–271.
Neumann, I., Brown, S., Smedley, P.L., Besien, T., 2003. Baseline Report Series: 7. The Great and the Inferior Oolite of the Cotswold district. British Geological Survey Commissioned Report CR/03/202N.
Noh, S.H., 2021. Analysis of gradient vanishing of RNNs and performance comparison. Information 12 (11), 442.
Oberhelman, A., Martin, J.B., Flint, M.K., 2024. Sources of limestone dissolution from surface water-groundwater interaction in the carbonate critical zone. Chem. Geol. 662, 122229.
Oubagaranadin, J.U.K., Sathyamurthy, N., Murthy, Z.V.P., 2007. Evaluation of fuller's earth for the adsorption of mercury from aqueous solutions: a comparative study with activated carbon. J. Hazard. Mater. 142 (1–2), 165–174.
Perzan, Z., Osterman, G., Maher, K., 2023. Controls on flood managed aquifer recharge through a heterogeneous vadose zone: hydrologic modeling at a site characterized with surface geophysics. Hydrol. Earth Syst. Sci. 27 (5), 969–990.
Shand, P., Tyler-Whittle, R., Besien, T., Lawrence, A.R., Lewis, O.H., 2003a. Baseline Report Series: 6. The Chalk of the Colne and Lee river catchments. British Geological Survey Commissioned Report CR/03/069N.
Shand, P., Cobbing, J.E., Tyler-Whittle, R., Tooth, A., Lancaster, A., 2003b. Baseline Report Series: 9. The Lower Greensand of southern England. British Geological Survey Commissioned Report CR/03/273C.
Shen, C., 2018. A transdisciplinary review of deep learning research and its relevance for water resources scientists. Water Resources Research 54 (11), 8558–8593.
Smedley, P.L., Griffiths, K.J., Tyler-Whittle, R., 2003. Baseline Report Series: 5. The Chalk of north Downs, Kent and east Surrey.
Tabari, H., 2020. Climate change impact on flood and extreme precipitation increases with water availability. Sci. Rep. 10 (1), 13768.
Tellman, B., Sullivan, J.A., Kuhn, C., Kettner, A.J., Doyle, C.S., Brakenridge, G.R., Erickson, T.A., Slayback, D.A., 2021. Satellite imaging reveals increased proportion of population exposed to floods. Nature 596 (7870), 80–86.
Tepetidis, N., Koutsoyiannis, D., Iliopoulou, T. and Dimitriadis, P., 2024. Investigating the performance of the informer model for streamflow forecasting.
Thakkar, A., Lohiya, R., 2022. A survey on intrusion detection system: feature selection, model, performance measures, application perspective, challenges, and future research directions. Artif. Intell. Rev. 55 (1), 453–563.
Wang, N., Zhao, X., 2023. Enformer: Encoder-based sparse periodic self-attention time-series forecasting. IEEE Access 11, 112004–112014.
Wang, W., Zhao, Y., Tu, Y., Dong, R., Ma, Q., Liu, C., 2023. Research on parameter regionalization of distributed hydrological model based on machine learning. Water 15 (3), 518.
Weytjens, H., De Weerdt, J., 2021. Creating unbiased public benchmark datasets with data leakage prevention for predictive process monitoring. In: International Conference on Business Process Management. Springer International Publishing, Cham, pp. 18–29. , September.
Yuan, F., Fan, C., Farahmand, H., Coleman, N., Esmalian, A., Lee, C.C., Patrascu, F.I., Zhang, C., Dong, S. and Mostafavi, A., 2022. Smart flood resilience: harnessing community-scale big data for predictive flood risk monitoring, rapid impact assessment, and situational awareness. Environ. Res. Infrastruct. Sustain., 2(2), p.025006.
Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H. and Zhang, W., 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).