https://doi.org/10.5194/egusphere-2026-3267
Preprint. Discussion started: 19 June 2026

Exploring the generalisation ability and interpretability of Long Short-Term Memory (LSTM) networks for large-sample groundwater level predictions

Qidong Fang[1], Mostaquimur Rahman[1], Thorsten Wagener[2], Francesca Pianosi[1]

[1]School of Civil, Aerospace and Design Engineering, University of Bristol, Bristol, BS8 1US, UK
[2]Institute of Environmental Science and Geography, University of Potsdam, Potsdam, 14476, Germany

Correspondence to: Qidong Fang (Qidong.fang@bristol.ac.uk)

Abstract. Deep Learning (DL) models, particularly Long Short-Term Memory (LSTM) networks, have shown similar or even superior performance to process-based models in estimating streamflow particularly at ungauged locations. However, their ability to extrapolate groundwater levels across time and space is less understood, as the number of studies addressing this issue is so far relatively limited. Here, we exploit the unique availability of a large-sample dataset of groundwater level observations across England to contribute to filling this gap. We configured two LSTM model variants: one using static environmental attributes (LSTM_ENV) and one using random integers as unique identifiers of places (LSTM_RND). Both models were trained using data from 636 stations over the period 1971-2014 and tested over 2015-2019 at both the training stations (in-sample test) and at 341 unseen stations (out-of-sample). Our results indicate that the two configurations achieved comparable performance in in-sample test, but their performances significantly diverge at unseen stations. To put the LSTM models' performance into context, we also compared them to the performance of a process-based surface-groundwater model at 124 unseen stations. We found that both models effectively capture temporal fluctuations but struggle to accurately reproduce the mean and variability of the water table depth. This systematic bias frequently resulted in negative NSE values despite high temporal correlation, suggesting that evaluating LSTM performance using NSE solely can be misleading. We also found that the LSTM_ENV model performs better at stations characterised by higher specific yield and transmissivity, and that it mostly uses meteorological input features (e.g. precipitation) and topographic features (e.g. elevation and height above nearest drainage) to make predictions at unseen stations. These findings highlight the potential of LSTMs for regional groundwater level predictions and the value of interpretability tools for understanding how such models achieve their performance and whether the environmental features used are informative.

Short Summary. It is unclear whether deep learning models can predict groundwater level at places without measurements using attributes of the places. Our deep learning model captured temporal variation well, especially in more responsive aquifers, similarly to a process-based model. Interpretation tools showed that meteorological and environmental information at places helped predictions at unseen wells. We highlight the potential of deep learning models for regional groundwater level predictions.

1 Introduction

Groundwater accounts for 99% of the world's liquid freshwater resource, serving as the primary source for nearly half of irrigated agricultural water use and the drinking water supply for billions of people (Siebert et al., 2010; Margat and van der Gun, 2013). Satellite data indicate that groundwater depletion has been occurring in several regions globally (Rodell et al., 2018; Rodell et al., 2009; Liu et al., 2022) and observation-based analyses show that groundwater levels declined rapidly, exceeding 0.5 m year^{-1}, in many parts of the world during the first two decades of the 21st century (Jasechko et al., 2024). Climate change may further intensify groundwater stress by increasing the frequency and severity of hydroclimatic extremes, while also increasing the human water demand in many regions (Green et al., 2011). These pressures highlight the critical

https://doi.org/10.5194/egusphere-2026-3267
Preprint. Discussion started: 19 June 2026

importance of effective groundwater management across scales, for which reliable groundwater level or depth prediction models serve as indispensable tools (Condon et al., 2021).

Historically, the simulation of groundwater dynamics generally falls into two categories. The first focuses on temporal reconstruction at gauged locations, where models aim to simulate fluctuations during periods not covered by monitoring records. This is typically addressed by developing single-station models (build one model for an individual station), such as conceptual (Rushton, 2004; Mackay et al., 2014) or empirical models (Rajaee et al., 2019; Tao et al., 2022; Wunsch et al., 2022). While these models can effectively capture local temporal variability, they cannot be generalised to ungauged areas. The second aims to simulate groundwater level fluctuations at locations where measurements are not available, traditionally the domain of distributed physically-based models (de Graaf et al., 2015; de Graaf et al., 2017; Yang et al., 2023; Bianchi et al., 2024). Nevertheless, developing large-scale physically-based models is still challenging: 1) hydrogeological data required for parameterisation and evaluation rarely cover the entire domain (Gleeson et al., 2021); 2) the computational cost of large-scale groundwater models can be prohibitively high (Reinecke et al., 2019).

In response to these limitations, multi-station Deep Learning (DL) models have emerged as a promising alternative. The underlying premise is that a multi-station DL model can learn transferable relationships from observations collected across multiple locations, thereby enabling generalisation across both temporal and spatial domains. Such a model was first introduced to hydrology by the work of Kratzert et al. (2018, 2019a, 2019b) for streamflow predictions. They trained multi-basin Long Short-Term Memory (LSTM) neural networks using dynamic meteorological forcings together with static catchment attributes from 531 catchments across the US. They found that the multi-basin LSTM models outperformed individually calibrated hydrological models in predicting streamflow at both gauged and ungauged basins (Kratzert et al., 2018, 2019a, 2019b).

Motivated by these advances in streamflow modelling, several studies have explored the application of multi-station LSTM models for groundwater level modelling, although such studies remain a minority. To the best of the authors' knowledge, only three studies have investigated multi-station LSTM modelling for groundwater applications, covering 76 stations in Northern France (Chidepudi et al., 2025), 108 stations in Germany (Heudorfer et al., 2024), and 1,800 stations across nine countries/regions (Nolte et al., 2025). Chidepudi et al. (2025) reported that LSTM models performed better at stations primarily influenced by annual cycles than at stations influenced by multi-annual cycles. Heudorfer et al. (2024) evaluated the performance of LSTM models using different static input configurations, including: 1) environmental attributes, such as land cover, mean annual average temperature, soil and aquifer types, distance to stream, etc., 2) random integers used as static features; and 3) without static inputs (using dynamic meteorological forcings only, i.e., precipitation, temperature and relative humidity). They found that using static features on top of dynamic ones improved performance, but also that the LSTM model using random integers performed similarly to the LSTM model fed with physically meaningful environmental features. Similar results were found by Chidepudi et al. (2025) and Nolte et al. (2025).

Notably, Heudorfer et al. (2024) also investigated spatial generalisation, i.e. the model's ability to make predictions at stations unseen during training. Again, they found that LSTM models using random integer features performed comparably to those employing meaningful environmental attributes. Furthermore, the models using only dynamic meteorological forcings even outperformed all other configurations at unseen stations. These findings imply that the LSTM model utilises static features primarily as 'unique identifiers' to memorise local station behaviours during training rather than deriving the generalisable hydrological insights required for spatial extrapolation. Similar experiments were conducted for streamflow prediction by Heudorfer et al. (2025), who again found that the predictive skill of LSTM models was dominated by meteorological signatures rather than by an effective use of physical catchment characteristics. Overall, these findings suggest that the spatiotemporal generalisation ability of multi-station DL models remains insufficiently understood. In particular, it remains unclear to what extent such models extract hydrologically meaningful information from environmental attributes that supports spatial transferability (Heudorfer et al. 2024, 2025; Baste et al., 2025).

https://doi.org/10.5194/egusphere-2026-3267
Preprint. Discussion started: 19 June 2026

Beyond their predictive capabilities, recent advances in post-hoc explainable AI (XAI) techniques have enabled improved interpretation of deep learning "black-box" models (Jiang et al., 2024; Slater et al., 2025). In hydrology, XAI techniques provide opportunities to better understand how deep learning models relate hydrometeorological and environmental inputs to simulated system responses (Jiang et al., 2024). Broadly, these techniques can address two questions. First, they can be used to examine how changes in feature values affect the direction and magnitude of model predictions (Molnar et al., 2022). This question can be addressed using methods such as Individual Conditional Expectation (ICE, Goldstein et al., 2015), local interpretable model-agnostic explanations (LIME, Ribeiro et al., 2016), and SHapley Additive exPlanations (SHAP, Lundberg and Lee, 2017). For example, Jung et al. (2024) used aggregated ICE analyses to identify an exponential increase in groundwater recharge rates with increasing long-term precipitation. Second, XAI techniques can be used to examine how features contribute to model performance (Molnar et al., 2022). This question can be addressed using methods such as Permutation Feature Importance analysis (PFI, Breiman, 2001; Fisher et al., 2019). By comparing features importance rankings, we can assess whether specific features used are informative. For example, in groundwater modelling, a low importance ranking for a hydrogeological attribute such as transmissivity may indicate that the available transmissivity data provide limited additional predictive information to the model. However, this does not necessarily imply that transmissivity is physically unimportant for groundwater dynamics (Freeze and Cherry, 1979; Rahman et al., 2023). Instead, it may suggest that the available representation of transmissivity is too coarse or uncertain. This can motivate further investigation into whether finer-resolution datasets, alternative spatial aggregation schemes, or different feature representations improve model performance (Tarasova et al., 2024).

In England, groundwater provides roughly a third of public water supplies, and over 75% in the densely populated and water-stressed Thames and Southern regions (BGS, 2024). A newly released large-sample groundwater dataset by the Environment Agency of England, comprising more than 200,000 daily and 200 million sub-daily sampling observations for over 3,400 wells, offers a unique opportunity to evaluate the generalisation ability of multi-station LSTM models in time and space. In this study, we want to investigate the following questions:

1. How well can a multi-station LSTM model simulate the groundwater variability across England?
2. Does the LSTM model perform better in certain locations, and if so, why?
3. Does the LSTM model actually learn from environmental features to extrapolate to unseen locations, and if so, which feature does the model use to make the prediction?

To address these questions, we first trained an LSTM model using dynamic meteorological forcings together with static environmental attributes. We also repeated the experiment designed by Heudorfer et al. (2024) of replacing the environmental features with the same number of random integers, to test whether LSTM models actually extract useful insights from the environmental features. Besides Nash-Sutcliffe Efficiency (NSE) and Kling-Gupta efficiency (KGE), which are the performance metrics used in other studies (Heudorfer et al., 2024; Chidepudi et al., 2025; Nolte et al., 2025), we also examined the KGE individual components, i.e. correlation, mean ratio and standard deviation ratio, to get a more in-depth understanding of the LSTM performance. We used the national-scale mechanistic DECIPHeR-GW model (Dynamic fluxEs and ConnectIvity for Predictions of HydRology and GroundWater, Zheng et al., 2025) for benchmarking the LSTM model. To our knowledge, this is the first comparison between DL and mechanistic models for groundwater modelling and will enable us to put the LSTM model performance into the context of what other modelling approaches can achieve. Lastly, we built a Classification Tree to see where the model performs well and used the Permutation Feature Importance analysis to investigate which input features influenced most to the LSTM performances.

https://doi.org/10.5194/egusphere-2026-3267
Preprint. Discussion started: 19 June 2026

2 Data

2.1 Groundwater depth timeseries

Groundwater depth observations from 2902 wells across England were obtained from Fang et al. (2025). This dataset is a polished version of the groundwater levels dataset released by the Environment Agency (EA) of England and Wales around 2023 (Environment Agency, 2023). In a previous study (Fang et al., 2026), we found that about half of the 2902 wells in this dataset experienced either a long-term trend or a sudden change. In this study, we focus on the 1641 stations showing neither a trend nor a sudden change, to investigate whether a DL model can learn the seasonal and inter-annual groundwater variation patterns rather than long-term trends or sudden changes. We also filtered out those wells that are potentially located in confined aquifers because the groundwater dynamic mechanisms in confined and unconfined aquifers can differ substantially. In the polished dataset (Fang et al., 2025), each station is linked to one of the 11 principal aquifers in England based on borehole depth, the lowest recorded groundwater level in the timeseries, and the elevation of the aquifer base (Further details of the aquifers assignment procedure are provided in Supplementary Material S3 of Fang et al. 2026). For example, in the east of England, the Crag aquifer overlies the Chalk aquifer. In this setting, the Crag aquifer can be assumed to be unconfined, whereas the underlying Chalk aquifer is likely to be confined. Therefore, wells in this area assigned to the Chalk aquifer were excluded from the present analysis. While this approach is a first step towards excluding wells in confined aquifers, we cannot fully guarantee that all the remaining wells are in unconfined aquifers, as apart from the 11 principal aquifers we analysed, there are many minor aquifers with no digital distribution information currently available (Jones et al., 2000). After this further selection, the number of wells is reduced to 1384.

We then examined the records distribution across time and stations. The groundwater dataset is highly imbalanced: about 75% of observations were collected after 2000, and about 75% of observations came from 20% of the wells (Fig. S1). We defined the period from 1 January 1971 to 31 December 2014 as the training period (~80% of total observations) and the period from 1 January 2015 to 31 December 2019 as the testing period. We then used the following criteria to further select stations based on data availability for training and testing:

1. Data quality: We excluded stations with negative mean water table depth. In fact, although groundwater table can be higher than the land surface in some cases, a large number of negative water table records are suspect. We also excluded stations with very low variability (Standard deviation of water table depth ≤ 0.1 m) as they exhibit no significant dynamics.
2. Data availability: We excluded stations that do not meet one (or more) of the three criteria: a) maximum gap between consecutive observations < 3 years. b) the mean annual data coverage ≥ 7 months and the intra-annual distribution averaged ≥ 3.6 seasons per year. c) the training period spans > 10 years.

We found 636 stations where data meet all the above criteria in both training period (1971-2014, for a total of 1,072,081 observations) and testing period (2015-2019, for a total of 310,241 observations). These stations will be used for training the LSTM model and for in-sample (IS) testing, i.e. testing the model's ability to extrapolate temporal patterns at the same locations that were used for the model training. To make the most use of the data, we will also use the 341 stations where the data only met the criteria in the testing period for spatially out-of-sample (OOS) testing, i.e. testing the model at locations unseen during training (total of 152,963 observations). A schematic of selecting groundwater stations and creating train/test datasets is shown in Fig. S3.

A recent study has indicated that LSTM performance in groundwater modelling is sensitive to the dominant fluctuation period of the timeseries (Chidepudi et al., 2025). We therefore applied wavelet transform analysis, following Baulon et al. (2022), to identify the dominant fluctuation cycle at each well for subsequent analyses. Specifically, each timeseries was classified according to whether its dominant fluctuation cycle is annual, multi-annual, or decadal.

https://doi.org/10.5194/egusphere-2026-3267
Preprint. Discussion started: 19 June 2026

Before conducting the wave analysis, the groundwater-level timeseries were resampled to a monthly basis, then the missing values were filled using linear interpolation. Because the resulting classification was used only for interpretation and was not used as an input to model training, the wavelet analysis was conducted over the full observation period from 1 January 1971 to 31 December 2019, without separating the data into training and testing periods. Using wavelet decomposition, each original timeseries was separated into a set of detail components and a final approximation component. The relative contributions of these components were then used to determine whether groundwater level dynamics at each well were primarily influenced by annual, multi-annual, or decadal oscillations, as illustrated in Fig. S2.

2.2 Meteorological time series

Groundwater recharge is a primary control on water table depth, and its magnitude and timing are strongly influenced by meteorological forcings. We therefore obtained a range of meteorological variables (namely: precipitation, air temperature, daily temperature range, specific humidity, wind speed, air pressure, downward longwave and shortwave radiation, and potential evapotranspiration) from the CHESS-met (Robinson et al., 2023a) and the CHESS-PE (Robinson et al., 2023b) datasets. Both datasets provide daily gridded data at 1km spatial resolution for the period from 1 January 1961 to 31 December 2019. A summary of meteorological variables is provided in Table A1.

2.3 Environmental attributes

2.3.1 Topography

Topography affects groundwater distribution and flow by shaping hydraulic gradients and drainage conditions. In low-permeability regions, the water table is often shallower and closely aligns with the local topography. Conversely, in regions with higher permeability, the water table is deeper and follows the regional topography (Gnann et al., 2025). To represent the effects of topography on water table depth variability, we used three attributes: elevation, slope, and height above nearest drainage (HAND). The surface elevation of each well was extracted from the 10m LIDAR Composite Digital Terrain Model (DTM) produced by the Environment Agency in 2022. Slope was then calculated from this dataset using QGIS. HAND is strongly associated with static water table depth, especially in steep terrain (Nobre et al., 2011; Janssen et al., 2025). We therefore extracted HAND values from the global HAND dataset at 3 arc-seconds resolution (Yamazaki et al., 2019).

2.3.2 Soil data

Soil properties influence the partitioning of precipitation into infiltration, runoff, and evapotranspiration, and therefore affect the magnitude and timing of recharge reaching the water table. We obtained clay, silt, and sand fractions, together with bulk density, for both topsoil (0 - 30 cm) and subsoil (30 - 70 cm), from the European Soil Database Derived data. In total, eight soil attributes were used in our study.

2.3.3 Hydrogeology

Hydrogeological properties control groundwater storage and flow through their influence on aquifer permeability, storage capacity, and subsurface connectivity. Geological structures, such as faults and fractures, can also affect groundwater flow pathways and hydraulic gradients. We used three hydrogeological attributes: transmissivity, specific yield, and distance from each well to the nearest fault. Spatial estimates of transmissivity and specific yield across England and Wales were taken from Rahman et al. (2023), who derived these attributes from the 1:625k geological dataset provided by the British Geological Survey (BGS). The distance from each well to the nearest fault was calculated from the same BGS 1:625k geological dataset using QGIS.

https://doi.org/10.5194/egusphere-2026-3267
Preprint. Discussion started: 19 June 2026

2.3.4 Water management

Water management can influence both rising and declining groundwater levels through abstraction, irrigation, and other anthropogenic modifications to the groundwater system (Fang et al., 2026; MacAllister et al., 2022). We used three water management related attributes: population density, irrigation density, and mean groundwater abstraction. Population density and irrigation density were taken from Fang et al. (2026). Mean groundwater abstraction for the period 1999-2014 was taken from Rameshwaran et al. (2025), who provided gridded monthly estimates of actual groundwater abstractions for England from January 1999 to December 2014, based on data from the Environment Agency (EA) of England and Wales. Due to the temporal coverage and resolution of the abstraction dataset, we did not use groundwater abstraction as a dynamic input to the LSTM network. Instead, we used mean groundwater abstraction at a 1km spatial resolution as a static attribute for each well. A summary of all environmental attributes is provided in Table A2.

3 Methods

3.1 LSTM structures

A Long Short-Term Memory (LSTM) network is a type of recurrent neural network designed for modelling sequential data (Hochreiter and Schmidhuber, 1997). Its core advantage lies in its ability to learn both short-term and long-term temporal dependencies through a memory cell and a set of gating functions that regulate information flow. A detailed description of the LSTM architecture is not repeated here; instead, the reader is referred to Hochreiter and Schmidhuber (1997) and Kratzert et al. (2019a, 2019b). In this study, the LSTM network was trained with observations from 636 stations (so-called multi-station LSTM) and to make the predictions at a daily resolution.

Inputs for multi-station LSTM include dynamic variables (i.e. varying over time) and static variables (i.e. constant over time). The dynamic variables consisted of the nine meteorological forcings (precipitation, air temperature, daily temperature range, specific humidity, wind speed, air pressure, downward longwave and shortwave radiation, and potential evapotranspiration) described in Sect. 2.2. Based on the presence of multi-annual variability in the groundwater records (Fig. S3b, c), we decided that the temporal window of these dynamic input variables should span back 10 years before the target day. This duration was chosen to balance the need for long-term memory with the practical constraints of data availability. However, as the training dataset comprises over one million target (groundwater level) observations, the computational cost would be extremely large if the 10-year-long record of each dynamic input variable were given at daily resolution (i.e. 3,650 time steps). We therefore aggregated the dynamic inputs into a combination of monthly and multi-month time steps. For the year immediately preceding the target day, meteorological variables were aggregated over 30-day intervals. For the preceding nine years, variables were aggregated over 270-day intervals. This results in 24 input time steps in total, comprising 12 time steps at each temporal resolution (Fig. 1c)

Static variables were time-invariant and had the same value for all target observations at a given station. We first trained an LSTM model using 17 environmental features described in Sect. 2.3 as static inputs (LSTM_ENV). We then trained a second LSTM model in which these 17 environmental features were replaced by the same number of random integers in the range [0, 9] (LSTM_RND). This experiment follows the rationale of Heudorfer et al. (2024) and was designed to test whether the LSTM model uses static inputs primarily as unique identifiers of place or whether it extracts meaningful and transferable information from environmental attributes. Apart from the different static input settings, both models shared the same architecture. The network consisted of one LSTM layer with 256 hidden states and a fully connected output layer. A dropout rate of 0.4 was applied to reduce the overfitting during training (Srivastava et al., 2014). The main hyperparameters, including learning rate, hidden state size, dropout rate, and number of LSTM layers, were tuned through grid search using 4-fold cross-validation. This tuning procedure followed the methodology described by Kratzert et al. (2019a) for multi-basin streamflow modelling.

https://doi.org/10.5194/egusphere-2026-3267
Preprint. Discussion started: 19 June 2026

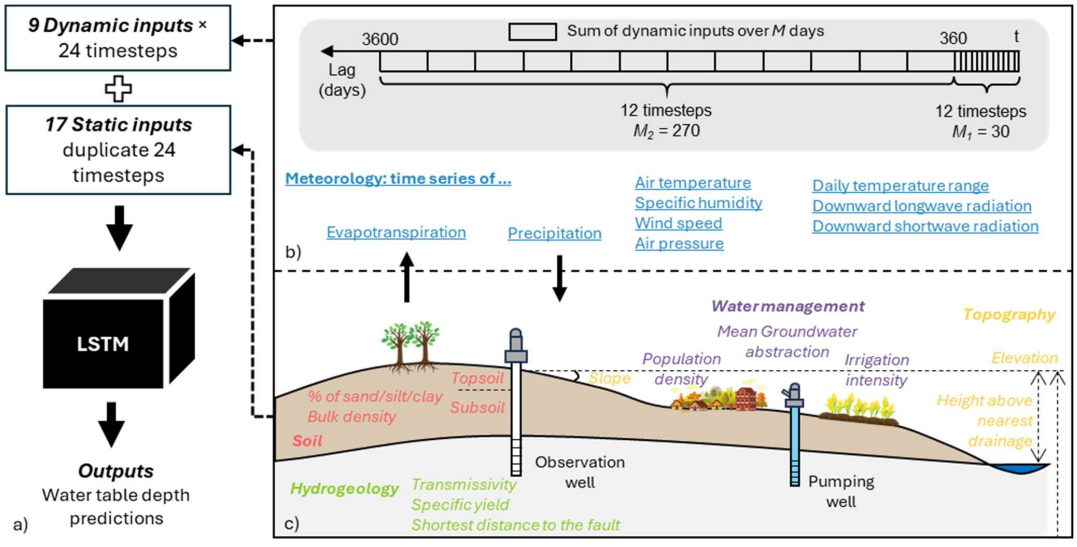


Fig. 1. Schematic of LSTM architecture and inputs. (a) LSTM model fed with 9 dynamic inputs and 17 static inputs. (b) Dynamic inputs are timeseries of meteorological variables (blue, underlined font). They are aggregated into a mix of monthly resolution (i.e. summed on a 30-day basis) and multi-month resolution (i.e. summed on a 270-day basis). (c) Static inputs (*italic font*) are classified into four categories: topography (yellow), soil (red), hydrogeology (green), and water management (purple).

3.2 Model training and testing

As illustrated in Sect. 2.1, the LSTM models are trained on 636 stations over the period 1971-2014 and tested on the same stations over the period 2015-2019 (in-sample testing, IS) and on 341 additional "unseen" stations over the period 2015-2019 (out-of-sample testing, OOS). The IS test tells us about the model's ability at temporal generalisation, the OOS test about both spatial and temporal generalisation ability.

For model training, we used station-averaged Nash–Sutcliffe Efficiency (NSE*, Nash and Sutcliffe, 1970) as the loss function, following Kratzert et al. (2019a). The definition of NSE* is the average of the NSE calculated at each station:

$$NSE^* = \frac{1}{M}\sum_{m=1}^{M}\left[\frac{\sum_{n=1}^{N}(y_n-\widehat{y_n})^2}{\sum_{n=1}^{N}(s(m)+\epsilon)^2}\right], \quad (1)$$

where M is the total number of groundwater stations $(m = 1, 2, \dots M)$; N is the number of observations in the station m $(n = 1, 2, \dots N)$; y_n is the observed value, $\widehat{y_n}$ is the predicted value, $s(m)$ is the standard deviation of observations at the station m. ϵ is a small constant added to the denominator to prevent numerical instability when $s(m)$ approaches zero, set as 0.1 following Kratzert et al. (2019a). Adopting NSE* ensures that each individual station contributes equally to the model optimisation, even when the number of observations across stations is highly unbalanced.

Given the inherent stochasticity in the network initialisation and optimisation during training, model outputs are subject to uncertainty. Kratzert et al. (2019a) showed that part of this uncertainty can be reduced through ensemble modelling, in which multiple LSTM networks are trained independently using different random initialisations (seeds) and the final predictions are then obtained by averaging the outputs of the individually trained models. The ensemble mean provides a more robust estimate of model performance than a single ensemble member by mitigating the influence of local minima during the optimisation process. In this study, we trained 10 LSTM models with different seeds for both configurations (LSTM_ENV and LSTM_RND). In the result section, we only show the predictions and performance from averaging the predictions of those 10 models, called "10-ensemble mean". For completeness, we also provide a performance comparison between the 10-ensemble mean and one of the ensemble members in the supplementary materials (see Fig. S4).

https://doi.org/10.5194/egusphere-2026-3267
Preprint. Discussion started: 19 June 2026

For model evaluation, we used six performance metrics: NSE, Kling-Gupta efficiency (KGE, Gupta et al., 2009), the three components of KGE, namely Pearson correlation, mean ratio, and standard deviation ratio; and Spearman rank correlation (for comparison with Pearson's). NSE was included to facilitate comparison with previous groundwater LSTM studies that have reported this metric. The remaining metrics were used to provide a more nuanced and detailed picture of the model performance. The definitions of all performance metrics are provided in Table A3.

3.3 Mechanistic benchmark model

DECIPHeR-GW (Dynamic fluxEs and ConnectIvity for Predictions of HydRology and GroundWater, Zheng et al., 2025) is a surface-groundwater hydrological model that couples the hydrological response units (HRU)-based surface hydrology model DECIPHeR (Coxon et al., 2019) and the two-dimensional gridded groundwater model by Rahman et al. (2023). DECIPHeR-GW was implemented and evaluated across England and Wales by Zheng et al. (2025). The model was calibrated using streamflow observations from 669 CAMELS-GB catchments (Coxon et al., 2020) for the period 1980-2010 and evaluated using groundwater level observations from 1804 unseen stations for the period 2010-2020. The groundwater observations used by Zheng et al. (2025) were sourced from the same Environment Agency datasets used in this study (Environment Agency, 2023).

For benchmarking, we identified 124 common stations between 1804 stations used in Zheng et al. (2025) and 341 out-of-sample stations used in this study. We then calculated the same performance metrics for both models over the testing period from 2015 to 2019. DECIPHeR-GW was calibrated in two ways: a) catchment-by-catchment; and b) nationally consistent. For a fairer comparison, we used the simulations from a nationally consistent calibration, as our LSTM model was also trained within a 'nationally consistent' way.

3.4 Analysis using explainable AI

In this study, we applied two complementary explainable AI (XAI) techniques: 1) Classification and Regression Trees (CART, Breiman et al., 1984), to investigate where and under which conditions LSTM_ENV perform well or poorly, and 2) Permutation Feature Importance (PFI, Breiman, 2001; Fisher et al., 2019), to quantify the relative contribution of each input feature to model performance.

3.4.1 CART analysis

CART is categorised as an 'interpretable model' or a 'white-box' method because its decision rules can be directly inspected (Breiman et al., 1984). Here, we used CART to identify hierarchical relationships between station attributes and LSTM_ENV model performance categories. The CART used 28 candidate predictors, including the 17 environmental attributes described in 2.3, the long-term means of 9 meteorological variables described in Sect. 2.2 for the period 1961–2019, the dominant fluctuation class derived from wavelet analysis (Fig. S3), and the number of observations available at each station. The output variable was the performance category of the LSTM_ENV (10-ensemble mean) model.

To better distinguish the patterns, we focused on stations exhibiting distinct performance extremes: Pearson $r \leq 0.25$ (poorly predicted) and Pearson $r > 0.75$ (well predicted) for further analysis. However, poor model performance at some stations may reflect inconsistent local groundwater behaviour rather than a limitation of the LSTM model itself (See Fig. S4). We therefore excluded poorly predicted stations that showed weak correspondence with nearby well predicted stations in the same aquifer. This filtering step was intended to focus the CART analysis on contrasts between well predicted stations and poorly predicted stations that were, in principle, expected to be predictable from environmental attributes.

To identify stations with inconsistent local behaviour, we conducted a correlation analysis between water table depth timeseries from neighbouring stations. First, all timeseries were resampled to monthly resolutions, and missing values were filled using linear interpolation. Second, for each poorly predicted station (Pearson $r \leq 0.25$), we identified well-predicted stations (Pearson

https://doi.org/10.5194/egusphere-2026-3267
Preprint. Discussion started: 19 June 2026

$r > 0.75$) located in the same aquifer. Third, we calculated the correlation between the timeseries of poorly predicted stations and those of nearby well predicted stations within a 20km radius. To avoid confusion with the Pearson correlation used as a model performance metric, we denote this inter-station timeseries correlation as *Corr*. The 20 km radius threshold was selected empirically based on manual inspection of neighbouring station behaviour. If the maximum *Corr* between a poorly predicted station and all nearby well predicted stations was less than 0.6, the station was labelled as exhibiting 'inconsistent local behaviour' and excluded from the CART analysis. The $Corr < 0.6$ threshold was also selected empirically based on manual inspection. In total 15 poorly predicted stations were excluded on this basis.

3.4.2 Permutation Feature Importance (PFI) analysis

To quantify the relative contribution of each input feature (described in Sect. 2.2 and 2.3) to LSTM performance, we applied a Permutation Feature Importance analysis (PFI, Breiman, 2001; Fisher et al., 2019). PFI estimates feature importance by measuring the degradation in the model performance after the association between a given input feature and the target variable is deliberately disrupted, without retraining the model.

In the PFI analysis, each feature was permuted across all samples and time steps. For static attributes, this permutation disrupted the spatial association between station attributes and groundwater depth. For dynamic meteorological variables, permutation disrupted both temporal and spatial structure in the forcing data. For each permuted feature, predictions were generated using the 10 trained LSTM ensemble members, and the resulting ensemble-mean prediction was used to calculate performance. The random permutation procedure was repeated 10 times for each feature.

Feature importance was quantified using an importance ratio, defined as the ratio between the loss obtained after permuting the feature j and the loss obtained using the original unpermuted input data:

$$IR_j = \frac{e_{permuted,\,j}}{e_{original}}, \quad (2)$$

An IR_j value close to 1 indicates that permuting the feature j has a negligible effect on model performance, whereas larger values indicate greater performance degradation and therefore higher feature importance. The error function e was defined as the median absolute deviation from the optimal metric value across all stations:

$$e = median(|1 - Metric|), \quad (3)$$

Here, $Metric$ refers to the six performance metrics described in Sect. 3.3 (i.e. NSE, KGE, Pearson *r*, Spearman *r*, Standard deviation ratio α, and mean ratio β).

Because multicollinearity among input features can bias PFI rankings and obscure the interpretation of individual feature importance (Jiang et al., 2024), we also implemented a grouped PFI analysis. In this analysis, features were permuted simultaneously within predefined categories in Sect. 2.2 and 2.3, allowing the importance of related groups of predictors to be assessed jointly.

4 Results

4.1 Performance of LSTM models at in-sample and out-of-sample stations

First, we evaluated the ability of the LSTM models to predict groundwater depth dynamics at in-sample stations, i.e. the same stations where the model was trained. We find that the performances are similar when we use environmental features (LSTM_ENV, red dashed lines in Fig. 2) or random integers (LSTM_RND, purple dashed lines in Fig. 2) as static inputs to the LSTM model. For both configurations, about 60% of the total 636 in-sample stations achieved NSE > 0.5 on a validation period unseen during training, and about 80% of stations achieved KGE > 0.5. About 80% of stations achieved Pearson *r* and Spearman $r > 0.75$, and the mean ratio are close to 1 at most stations, implying that the temporal variation and mean of the water table depth timeseries are predicted well (See Fig. S7 for some examples of observed and predicted water table depth).

https://doi.org/10.5194/egusphere-2026-3267
Preprint. Discussion started: 19 June 2026

However, about 80% of stations had a Standard Deviation ratio < 1, indicating the amplitude of water table depth fluctuations was generally underestimated. Similar to Chidepudi et al. (2025), we also found that all LSTM models perform better at stations primarily influenced by annual cycles (Fig. S6a).

We then examined model performance at spatially out-of-sample stations, i.e. stations not seen during training. Across the 341 out-of-sample stations (solid lines in Fig. 2), both configurations exhibited substantial performance degradation across all metrics compared to the spatially in-sample evaluation (dashed lines). Notably, NSE values were predominantly negative for the out-of-sample test. Nevertheless, the LSTM model fed with environmental features as static inputs (LSTM_ENV, red solid lines) demonstrated superior performance. achieving positive KGE values at about 45% of the total 341 out-of-sample stations compared to about 30% for the LSTM fed with random integers (LSTM_RND, purple line).

Decomposition of the KGE metric indicates that this performance gap was primarily associated with correlation. About 40% of out-of-sample stations achieved correlation coefficients (both Pearson *r* and Spearman *r*) greater than 0.75 with LSTM_ENV, compared with about 20% for LSTM_RND. LSTM_ENV also performed better than LSTM_RND in reproducing the mean and fluctuation amplitude of the water table depth timeseries, although the differences were smaller. For LSTM_ENV, about 50% and 40% of stations fell within the range [0.5, 1.5] for the mean ratio and standard deviation ratio, respectively, compared with about 35% and 30% for LSTM_RND. Overall, spatiotemporal generalisation remained challenging, but LSTM_ENV showed better transferability than LSTM_RND, particularly in capturing temporal variability as measured by Pearson and Spearman correlation (See Fig. S8 for some examples of predicted and observed timeseries at selected stations). As in-sample evaluation, both models performed better at stations dominated by annual cycle (Fig. S6b).

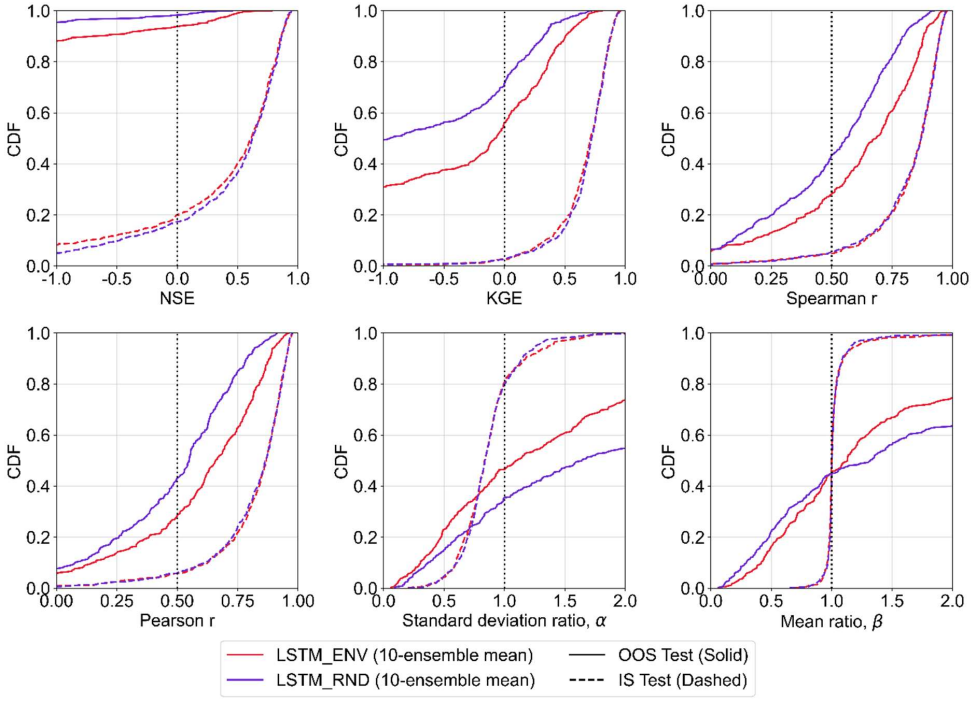


Fig. 2. Cumulative distribution functions (CDFs) of performance metrics for LSTM models that use environmental features (LSTM_ENV; red) and random integers (LSTM_RND; purple) as static inputs. Dashed lines show results for the 636 spatially in-sample stations, and solid lines show results for the 341 out-of-sample stations (i.e., stations not used for the LSTM training). Predictions were obtained by averaging predictions from 10 models trained with different random initialisations (seeds). For NSE, KGE, Spearman *r*, and Pearson *r*, superior performance is indicated by curves shifted towards the lower-right corner. For the Standard deviation ratio (α) and Mean ratio (β), superior performance is indicated by closer proximity to the vertical line at a ratio of 1.

https://doi.org/10.5194/egusphere-2026-3267
Preprint. Discussion started: 19 June 2026

4.2 Spatial patterns of performance

In this section, we investigate whether LSTM performance exhibits spatial patterns and whether out-of-sample stations with particularly good or poor performance share common characteristics. To this end, we constructed a CART model using 28 attributes (See Sect. 3.4 for details). The target variable was a binary performance class based on Pearson *r* of the LSTM_ENV (10-ensemble mean): poorly predicted, defined as Pearson $r \leq 0.25$ ($n = 32$ after excluding stations showing inconsistent local behaviour: See Sect. 3.4), and well predicted, defined as Pearson $r > 0.75$ ($n = 127$).

The classification tree reveals clear hydrogeological controls on model performance (Fig. 3a). Specific yield emerges as the primary splitting variable: stations located in aquifers with higher specific yield were predominantly associated with better LSTM performance (i.e. high Pearson *r*). Specifically, 99 out of 106 stations (93%) with specific yield greater than 0.012 have Pearson $r > 0.75$. In fact, higher specific yield was also generally associated with higher transmissivity, and 75 of these 99 well predicted stations were located in the Chalk aquifer (Fig. 3b), where transmissivity exceeded 500 $m^2 d^{-1}$. The Chalk aquifer is a major fractured carbonate aquifer in the UK, characterised by a high matrix porosity but where groundwater flow is predominantly fracture-dominated. In this region, well-developed fracture systems and high transmissivity are likely to promote relatively rapid groundwater level responses to recharge (Allen et al., 1997), which may favour more predictable temporal dynamics.

Among the 52 stations with specific yield less than and equal to 0.012, 28 stations (54%) located in aquifers still achieved Pearson $r > 0.75$. We do not discuss the subsequent split of this branch further because the poorly predicted stations in this group are spatially sparse across England (Fig. 3c), and we do not have an explanation for the next splitting variable, namely slope. Although the tree suggests that stations with steeper slopes were more likely to achieve Pearson $r > 0.75$, this pattern should be interpreted cautiously.

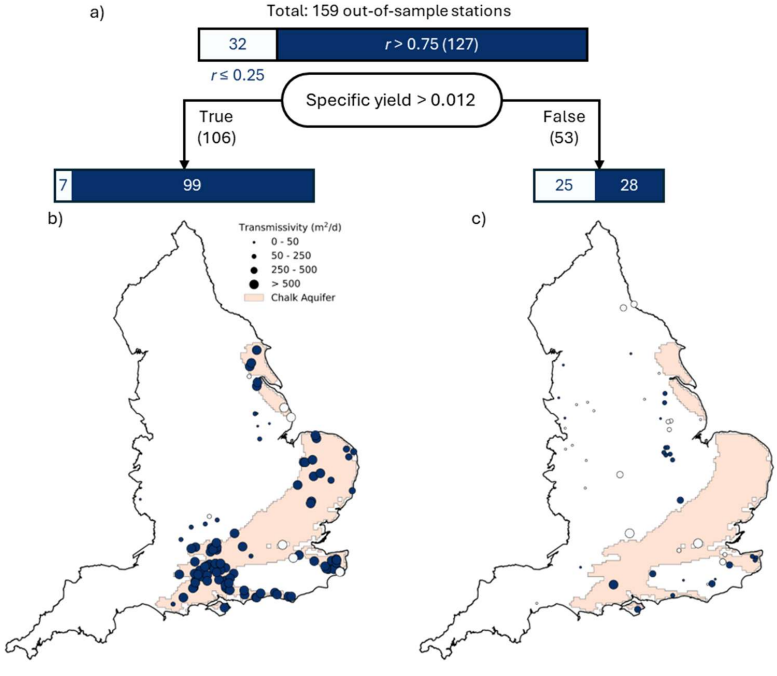


Fig. 3. Classification tree analysis of LSTM performance. (a) Classification tree constructed for 159 out-of-sample stations classified as either poorly predicted (Pearson $r <= 0.25$, n=32) or well predicted (Pearson $r > 0.75$, n=127) by LSTM_ENV (10-ensemble mean). The bottom panel shows the spatial distribution of two classes: (b) specific yield > 0.012 and (c) specific yield ≤ 0.012. Circle sizes indicates transmissivity. The Chalk aquifer is shown in the background.

https://doi.org/10.5194/egusphere-2026-3267
Preprint. Discussion started: 19 June 2026

4.3 Benchmarking the LSTM model with the mechanistic DECIPHeR-GW model

To put the LSTM results into context, we compared the model performance with that of the mechanistic surface-groundwater model DECIPHeR-GW (Zheng et al., 2025). To this end, we selected a subset of 124 out-of-sample stations common to both studies. Overall, the performance of LSTM_ENV (10-ensembles mean) is essentially the same as DECIPHeR-GW across all performance metrics (Fig. S10). Both models captured temporal variations in water table depth reasonably well, as shown by high correlation values. However, both showed weaker performance in reproducing the mean and variability of the water table depth timeseries, as reflected by poorer performance in the mean ratio and standard deviation ratio. These patterns are also consistent with the LSTM_ENV results presented in Sect. 4.1 for the full set of 341 spatially out-of-sample stations.

We further used Pearson correlation coefficient (r) to investigate whether there were systematic cases in which one model outperformed the other. Based on the relative performance, the stations were categorised into three classes (Fig. 4a, b): 1) Both models perform well (grey): about 41% of stations, where both models effectively capture temporal variations, with $r > 0.75$ for at least one model and the difference between models $\Delta r < 0.25$. 2) The LSTM_ENV model outperforms DECIPHeR-GW (red): about 11% of stations with a high LSTM_ENV performance ($r > 0.75$) and a substantial increase over DECIPHeR-GW ($\Delta r > 0.25$). 3) DECIPHeR-GW outperforms LSTM_ENV (yellow): about 7% of stations. We have also repeated this analysis using Spearman correlation coefficient as the indicator, which measures non-linear relationships between simulated and observed timeseries, and the conclusion stays the same (See Fig. S11).

Spatially, no notable spatial patterns emerged to indicate a systematic advantage of one model over the other in specific regions (Fig. 4b). However, both models showed high performance in the southeastern region of the Chalk aquifer, which is also consistent with the spatial pattern we found in Sect. 4.2. A representative example from a station in the Chalk aquifer is presented in Fig. 4c. It illustrates that both models excelled at simulating temporal correlation, but they both showed a systematic bias in the mean and variability of the water table depth timeseries.

https://doi.org/10.5194/egusphere-2026-3267
Preprint. Discussion started: 19 June 2026

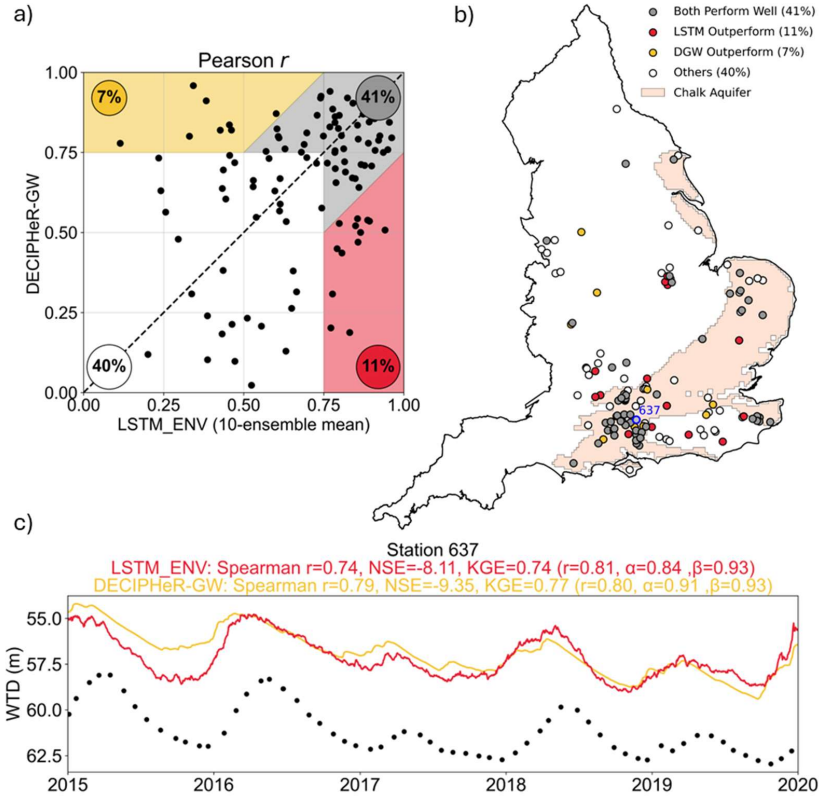


Fig. 4. Performance comparison between LSTM_ENV (red) and DECIPHeR-GW (yellow) across 124 common out-of-sample stations. (a) Comparison of Pearson correlation coefficient. (b) Spatial distribution of stations (circles) categorised by their relative Pearson *r*: grey indicates high performance for both models; yellow indicates that DECIPHeR-GW outperforms LSTM_ENV, red indicates that LSTM_ENV outperforms DECIPHeR-GW, and white represents all remaining cases. The extent of the Chalk aquifer is shown in the background. The blue circle highlights Station 1092, an example where both models capture temporal correlation effectively. (c) Observed and simulated water table depth (WTD) timeseries for Station 637. Black dots represent observations.

4.4 Feature importance of LSTM

To examine whether the LSTM_ENV model relied on meteorological variables or environmental attributes when making predictions to unseen locations, we applied Permutation Feature Importance (PFI, Breiman, 2001; Fisher et al., 2019) analysis to the training set, in-sample test set, and out-of-sample test set. Grouped PFI revealed a clear separation between the types of features supporting different aspects of model performance (Fig. 5a). Across all subsets (training, in-sample testing and out-of-sample testing), topographic attributes were consistently the most important feature group for NSE, KGE, the standard deviation ratio and the mean ratio. This indicates that the LSTM_ENV relies strongly on topographic information to infer the mean and variability of water table depth at unseen stations. In contrast, meteorological variables showed the highest grouped importance for Pearson *r* and Spearman *r*. This pattern was also consistent across three subsets, suggesting that dynamic meteorological forcings primarily support the model's ability to reproduce the temporal dynamics of water table depth.

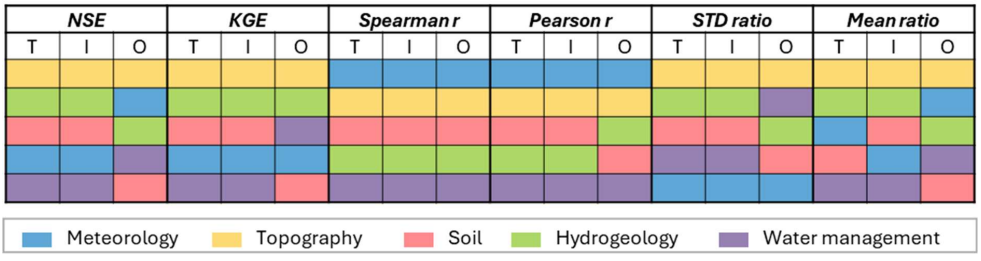
Individual PFI further showed that elevation (elev) and height above nearest drainage (hand) were the dominant individual features for all performance metrics across the three subsets (Fig. 5b, Fig. S13). Precipitation (P) was among the most important features for the correlation metrics. Air pressure (PSURF) also appeared as an important individual feature in several metrics. This is surprising and we explain it by noticing that long-term air pressure is strongly correlated with elevation (-0.96, see Fig. S12), so that its importance likely reflects its role as an elevation proxy rather than a direct meteorological control on

https://doi.org/10.5194/egusphere-2026-3267
Preprint. Discussion started: 19 June 2026

groundwater dynamics. Other individual meteorological variables had relatively low PFI ranking. This can be explained by the strong inter-variable correlation among the meteorological variables (Fig. S12) as when one meteorological variable is permuted at a time, the other correlated variables may retain similar seasonal information.

Hydrogeological attributes had secondary but consistent grouped importance, particularly for NSE, KGE, the standard deviation ratio and the mean ratio (Fig. 5a). They showed low grouped importance in terms of correlation metrics, yet relatively high individual feature importance, specifically for shortest distance to the fault (sdf, ranked 4th for both Pearson *r* and Spearman *r*) and transmissivity (ts, ranked 6th for Spearman *r* and 7th for Pearson *r*). Soil and water management attributes generally showed lower grouped importance, although variable such as population density (pop_dens, ranked from 4th to 6th across metrics except the mean ratio) contributed to most performance metrics (Fig. 5b). Possible explanations for this ranking result will be discussed in the next section.

a) Grouped feature importance across Train (T) / In-sample test (I) / Out-of-sample test (O) subsets



b) Individual feature importance on out-of-sample test subsets

NSE	KGE	Spearman r	Pearson r	STD ratio	Mean ratio
elev	elev	hand	hand	elev	elev
hand	hand	P	elev	hand	hand
sdf	sdf	elev	P	sdf	PSURF
PSURF	pop_dens	sdf	sdf	pop_dens	ts
ts	PSURF	pop_dens	pop_dens	PSURF	WIND
pop_dens	slope	ts	slope	slope	sdf
slope	ts	slope	ts	irri_int	DTR
irri_int	irri_int	PSURF	PSURF	ts	slope
sy	P	RLDS	irri_int	sy	HUSS
P	RLDS	irri_int	sy	TAS	t_sand
RLDS	sy	sy	RLDS	s_bd	RLDS
WIND	TAS	PET	RSDS	s_clay	PET
RSDS	RSDS	RSDS	PET	s_silt	RSDS
t_sand	PET	t_silt	t_silt	t_silt	s_sand
PET	s_clay	s_silt	s_bd	RLDS	P
s_bd	s_silt	s_bd	s_clay	PET	t_clay
t_clay	s_bd	s_clay	s_silt	P	s_bd
HUSS	t_silt	TAS	s_sand	WIND	t_bd
s_sand	WIND	s_sand	TAS	DTR	s_silt
TAS	s_sand	t_clay	gw_abs	RSDS	gw_abs
s_clay	t_clay	t_sand	t_clay	gw_abs	pop_dens
DTR	t_bd	gw_abs	t_sand	s_sand	sy
gw_abs	gw_abs	t_bd	t_bd	HUSS	TAS
s_silt	DTR	WIND	WIND	t_clay	s_clay
t_silt	t_sand	HUSS	HUSS	t_bd	t_silt
t_bd	HUSS	DTR	DTR	t_sand	irri_int

Abbr.	Feature name
P	Precipitation
TAS	Air temperature
DTR	Daily temperature range
HUSS	Specific humidity
WIND	Wind speed
PSURF	Air pressure
RLDS	Downward longwave radiation
RSDS	Downward shortwave radiation
PET	Potential evapotranspiration
elev	Elevation
slope	Slope
hand	Height above nearest drainage
t_sand	Percentage of sand in topsoil
t_silt	Percentage of silt in topsoil
t_clay	Percentage of clay in topsoil
t_bd	Bulk density in topsoil
s_sand	Percentage of sand in subsoil
s_silt	Percentage of silt in subsoil
s_clay	Percentage of clay in subsoil
s_bd	Bulk density in topsoil
sdf	Shortest distance to the fault
ts	Transmissivity
sy	Specific yield
pop_dens	Population density
irri_int	Irrigation intensity
gw_abs	Mean groundwater abstraction

Fig. 5. Grouped and individual permutation feature importance of the LSTM_ENV model. (a) Grouped feature importance for six model performance metrics across Train (T), In-sample test (I), and Out-of-sample test (O) subsets, ranked in descending order of importance. (b) Individual feature importance on out-of-sample test subsets. The left panel shows the relative importance of individual

https://doi.org/10.5194/egusphere-2026-3267
Preprint. Discussion started: 19 June 2026

features across six model performance metrics. The right panel provides the full features' name for each abbreviation. Detailed descriptions of features are available in Tables A1 and A2. Dynamic inputs are timeseries of meteorological variables (blue, UPPERCASE). Static inputs (lowercase) are classified into four categories: topography (yellow), soil (red), hydrogeology (green), and water management (purple).

5 Discussion

Overall, we found that the performances of LSTM_ENV and LSTM_RND models are comparable when the models are evaluated at in-sample stations over an unseen period. This suggest that the LSTM model primarily uses static inputs as 'unique identifiers' of places that help memorise local temporal dynamics. This is aligned with previous studies in hydrology who also found that LSTM models using location identifiers as static inputs show no significant difference with LSTM models fed with environmental attributes (Chidepudi et al., 2025; Heudorfer et al., 2024; Nolte et al., 2025; Li et al., 2022). We also found that both LSTM models achieved better performance at stations dominated by annual cycles, as previously shown by Chidepudi et al. (2025).

When tested at out-of-sample stations, we found that LSTM_ENV outperformed LSTM_RND, especially in capturing the temporal fluctuations of water table depth, as shown by a high correlation coefficient between predictions and observations. Diverging from other studies that employed multi-station LSTM models for groundwater modelling, we were able to compare the performance not only across different setups, but also with a mechanistic model (DECIPHeR-GW, Zheng et al., 2025), and found that LSTM_ENV yields similar performances to DECIPHeR-GW at 124 unseen stations.

The better performance of LSTM_ENV over LSTM_RND at unseen stations contradicts the findings of Heudorfer et al. (2024) who reported both setups (LSTM fed with environmental attributes or random integers) to have similar performance at out-of-sample stations. Our results suggest that, in the present application, LSTM_ENV was able to extract transferable information from the meteorological and environmental attributes for spatial extrapolation. This interpretation is supported by the consistency of the Permutation Feature Importance (PFI) patterns between the training and out-of-sample testing sets. Specifically, topographic attributes (i.e. elevation and height above nearest drainage) contributed mainly to representing spatial differences. This is hydrologically plausible because these variables encode drainage potential and relative proximity to surface-water drainage pathways, which are closely related to groundwater storage and water table position (Nobre et al., 2011; Fan et al., 2013; Janssen et al., 2025). Whereas dynamic meteorological forcings, especially precipitation, contributed mainly to capturing temporal groundwater dynamics (Fig. 5). For LSTM_RND, performance was good at some out-of-sample stations (Pearson $r > 0.75$ was achieved at about 15% stations, see Fig. 2). It appears that the LSTM_RND model learned some patterns useful for temporal and spatial extrapolation from dynamic meteorological variables, i.e. precipitation (P) and air pressure (PSURF, which is highly correlated with elevation, see Fig. S12). We have repeated the permutation feature importance for LSTM_RND to prove this (see Fig. S15).

We also found that while LSTM_ENV delivers a good performance in terms of correlation between observations and predictions, it rarely achieves a positive NSE at out-of-sample stations. We argued that this discrepancy does not necessarily indicate a failure of our LSTM model to learn regional hydrological patterns; rather, it highlights the limitation of the NSE as a robust performance metric in this context. Similar concerns increasingly supported by recent literature in streamflow modelling, which shows that aggregated performance metrics like NSE can mask compensating errors and offer limited diagnostic insight (do Nascimento et al., 2026; Williams, 2025). In our study, many stations capture temporal dynamics well (as shown in high r in Figs. 2, 3), yet yielded negative NSE score due to the bias between predicted and simulated mean water table depth (See Fig. S9 for some examples where r is higher but NSE is very negative because the mean water table depth is estimated by the LSTM at few meters above or below the observations). In large-scale groundwater modelling, such biases are often unavoidable due to large hydrogeological heterogeneity, a lack of localised parameters such as transmissivity, and the scale mismatch between point observations and, for example, coarse-resolution meteorological variables (Zheng et al., 2025; Rahman et al., 2023; Bianchi et al., 2024). In many applications, an error of few meters in mean water table may be acceptable, especially where the water table is deep and the relative error is small. In such cases, accurately capturing the temporal

https://doi.org/10.5194/egusphere-2026-3267
Preprint. Discussion started: 19 June 2026

fluctuation maybe more important than reproducing the absolute mean table depth. Therefore, we suggest that evaluating model performance solely through NSE may obscure the model's true ability, and using multiple performance metrics, such as KGE components (i.e., correlation, mean ratio, and standard deviation ratio), or other groundwater signatures (Heudorfer et al., 2019; Collenteur et al., 2025), can provide a better evaluation of the model performance in its different aspects.

Focusing on the Permutation Feature Importance (PFI) results for correlation metrics (i.e. Pearson *r* and Spearman *r*), the comparison between individual and grouped PFI suggests that the rankings were affected by collinearity within each feature group (Jiang et al., 2024). Soil attributes showed moderate-to-low grouped importance but generally low individual rankings. This can be partly explained by the strong inter-correlation among soil attributes (see Fig. S12). This suggests that transferable soil information was collectively useful but distributed across correlated attributes. It may also reflect a mismatch between attribute scale and hydrological process: soil attributes were taken from the 1km grid cell containing each well, whereas soil controls on recharge may operate over broader recharge areas (Fu et al., 2019). In contrast, hydrogeological and water management attributes showed lower grouped importance, but several individual attributes, including shortest distance to the fault, transmissivity, population density and irrigation intensity, still ranked highly (Fig. 5). These variables were less correlated within their respective groups and may therefore provide more distinct predictive information. Transmissivity and shortest distance to fault may provide transferable information on aquifer properties and groundwater flow types, helping the model distinguish between different aquifer behaviours (Allen et al., 1997; Yeh et al., 2016). Population density may act as a proxy for the combined effects of human activities in urban areas, such as changes in groundwater abstractions associate with industrialisation/de-industrialisation, leakages from water mains, and other urban water-management effects (Fang et al., 2026). Conversely, the lower rankings of specific yield and mean groundwater abstraction do not necessarily imply that storage properties or groundwater abstraction are unimportant. Indeed, specific yield may carry the transferable information already partly included in transmissivity (and/or shortest distance to fault), as we found that higher specific yield was also generally associated with higher transmissivity (Fig. 3). As for groundwater abstractions, a possible reason for their low importance ranking is that the temporal aggregation to annual mean hinders seasonal or event-scale pumping effects. In summary, feature importance analysis do not necessarily reveal new insights on physical processes but rather seem to highlight the limitations in the attributes used here, which are possibly not aggregated in a hydrological meaningful way (Tarasova et al., 2024).

The Permutation Feature Importance (PFI) analysis should be interpreted as a performance-based assessment of feature relevance under specific evaluation distributions, rather than as a complete description of the internal prediction mechanism of the LSTM models. As highlighted by Molnar et al. (2022), PFI computed on unseen test data is appropriate for assessing how features contribute to model generalisation performance. By contrast, methods such as SHapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) are more directly aimed at quantifying feature contributions to individual model predictions. Therefore, our results only showed which environmental and meteorological attributes are important for maintaining LSTM performance under spatial extrapolation, but they do not fully reveal how the LSTM internally combines these attributes to generate individual predictions. Overall, a more detailed interpretability analysis is therefore needed to diagnose the causes of poor model performance at specific stations and to assess whether low-ranking environmental features are genuinely uninformative or simple poorly represented for the relevant groundwater process scale. This remains a promising but challenging task for future work.

Lastly, the development of our LSTM models involved several iterative refinements. We compiled six lessons of 'dos and don'ts' that we learned from our modelling experience (Table 1). While some of these points are documented in existing literature, we reiterate them here to emphasise their importance for the broader hydrological modelling community. Most notably, for example, our initial assessment focused only on the model's performance on temporal extrapolation within the stations, which is common practice to evaluate models' performance (Chidepudi et al., 2025; Nolte et al., 2025). However, good performance in temporal prediction alone does not imply the model is able to generalise across space. Therefore, we

https://doi.org/10.5194/egusphere-2026-3267
Preprint. Discussion started: 19 June 2026

advocate that such multi-station LSTM (or more broadly Deep Learning) models intended for predictions across time and space must be tested on both in-sample and out-of-sample stations.

Table 1. Six lessons we learnt on dos and don'ts when applying LSTM models to groundwater predictions.

Stage	Checklist	Reasons
Data pre-processing	Remove groundwater level timeseries that show very little variation over time.	It is difficult for models to capture meaningful patterns or "dynamic" in data that is almost flat.
	Identify if the groundwater levels are dominated by annual or multi-year cycles.	These cycles reveal the 'memory' of the groundwater system; knowing the cycle length helps you decide how many years of meteorological forcing the model needs to see.
Model setups	Use multi-year meteorological forcing as dynamic inputs if there are stations dominated by multi-year cycle.	If the system has a long memory, the model needs a longer history of meteorological forcing to accurately predict current levels.
	Train an "ensemble" of multiple models with different initialisation instead of relying on just one LSTM model.	Different initialisations can lead to uncertainty in the model's outputs. Training multiple models can effectively reduce this uncertainty and improve the models' performance.
Model evaluation	Use multiple performance metrics, such as KGE decompositions (i.e., correlation, mean ratio, and standard deviation ratio) rather than only NSE.	Unlike NSE, KGE decomposition allows for the separate assessment of temporal correlation, bias and variability, providing a more detailed view of hydrological consistency.
	When testing model configurations (such as input selection, model architecture, training strategies, etc.), prioritise testing on out-of-sample stations rather than just spatially in-sample stations.	The ultimate goal of training such multi-station LSTM is spatiotemporal extrapolation, predicting at unseen locations and across different time periods. We are unable to evaluate the extent to which the model has learned useful information for spatial extrapolation from the data if we only focus on the model performance on spatially in-sample stations.

6 Conclusions

This study evaluated the spatiotemporal extrapolation ability of two LSTM configurations (using environmental attributes and random integers as static inputs, respectively) across 636 training stations and 341 spatially out-of-sample stations. While both models exhibited comparable skill in the in-sample test, the LSTM_ENV model demonstrated significantly superior performance in the out-of-sample test, particularly in capturing temporal dynamics at unseen stations. This result suggests that environmental attributes are beneficial for extrapolating in space.

A comparison between the LSTM and the mechanistic DECIPHeR-GW model revealed a shared strength in simulating temporal fluctuations, but a common limitation in reproducing the mean and variability of timeseries. This systematic bias frequently resulted in negative NSE values despite high temporal correlation. Evaluating LSTM performance using widely used efficiency metrics (i.e. NSE) can lead to highly misleading outcomes given large biases and small variances of residuals and observed variables. Disaggregating NSE into its constituent elements avoids this problem and provides more information about variability of error characteristics, i.e. enabling us to independently assess temporal correlation, magnitude-based errors, and broader hydrological signatures.

Hydrogeological analysis revealed that LSTM performance is significantly better in aquifers characterised by high specific yield and transmissivity. These physical properties appear to be intrinsically linked to rapid water table response to recharge. The consistency of grouped Permutation Feature Importance (PFI) patterns between training and out-of-sample evaluation suggests that the LSTM_ENV exploits transferable information from meteorological (e.g. precipitation) and environmental (e.g., elevation and height above nearest drainage) attributes when extrapolating to unseen locations. Lower-ranking individual attributes, such as soil attributes, specific yield and groundwater abstraction, may be explained by the strong inter-correlation within feature group or lack of hydrological meaningful aggregation or representation.

Overall, our results indicate that multi-station LSTM models have substantial potential for regional groundwater predictions, but that spatial generalisation remains to be improved. We highlight the value of interpretability tools for understanding how

https://doi.org/10.5194/egusphere-2026-3267
Preprint. Discussion started: 19 June 2026

such models achieve their performance and whether the environmental features used are informative. Further work is needed to determine how information encoded within Deep Learning models can be more robustly interpreted, so as to diagnose the causes of poor model performance at specific stations and to assess whether low-ranking environmental features are genuinely uninformative or simple poorly represented for the relevant groundwater process scale.

Appendix A

Table A1. Summary of meteorological attributes.

Class	Attribute	Description	Unit	Period	Temporal and spatial resolution	
Climatology	P	Precipitation (derived from the Met Office national database of observed precipitation)	mm day^{-1}	1961-2019	Daily	1 km
	TAS	Air temperature	K	1961-2019	Daily	1 km
	DTR	Daily temperature range	K	1961-2019	Daily	1 km
	HUSS	Specific humidity	kg kg^{-1}	1961-2019	Daily	1 km
	WIND	Wind speed	m s^{-2}	1961-2019	Daily	1 km
	PSURF	Air pressure	Pa	1961-2019	Daily	1 km
	RLDS	Downward longwave radiation	W m^{-2}	1961-2019	Daily	1 km
	RSDS	Downward shortwave radiation	W m^{-2}	1961-2019	Daily	1 km
	PET	Potential evapotranspiration for a well-watered grass (calculated using Penman-Monteith equation)	mm day^{-1}	1961-2019	Daily	1 km

https://doi.org/10.5194/egusphere-2026-3267
Preprint. Discussion started: 19 June 2026

Table A2. Summary of environmental attributes.

Class	Attribute	Description	Unit	Spatial resolution
Topography	elev	Station elevation in meters above ordnance datum	mAOD	10 m
	slope	Slope generated in QGIS	°	10 m
	hand	Height above nearest drainage	m	3 arc-seconds
Soil	t_sand	Percentage of sand in topsoil (0 - 30 cm)	%	1 km
	t_silt	Percentage of silt in topsoil (0 - 30 cm)	%	1 km
	t_clay	Percentage of clay in topsoil (0 - 30 cm)	%	1 km
	t_bd	Bulk density in topsoil (0 - 30 cm)	g cm^{-3}	1 km
	s_sand	Percentage of sand in subsoil (30 - 100 cm)	%	1 km
	s_silt	Percentage of silt in subsoil (30 - 100 cm)	%	1 km
	s_clay	Percentage of clay in subsoil (30 - 100 cm)	%	1 km
	s_bd	Bulk density in topsoil (30 - 100 cm)	g cm^{-3}	1 km
Hydrogeology	sdf	Shortest distance to the fault	km	1:625,000
	ts	Estimated transmissivity	m^2 d^{-1}	1:625,000
	sy	Estimated specific yield	-	1:625,000
Water management	pop_dens	Population density estimated on Census 2021	km^{-2}	MSOAs*
	irri_int	Estimated irrigation intensity in 2010	m^3 km^{-2}	About County scale
	gw_abs	Mean actual groundwater abstraction over 1999-2014	m^3 $month^{-1}$	1 km

* Middle layer Super Output Areas (MSOAs) comprise between 2,000 and 6,000 households and have a usually resident population between 5,000 and 15,000 persons. Population density is then calculated within a 5 km radius of each station.

Table A3. Definition of performance metrics.

Performance metrics	Equation	Note
NSE	$NSE = 1 - \frac{\sum_{i=1}^{N}(O_i - P_i)^2}{\sum_{i=1}^{N}(O_i - \bar{O})^2}$	Range: (-∞, 1], unit less, larger is better.
KGE	$KGE = 1 - \sqrt{(r_p - 1)^2 + (\alpha - 1)^2 + (\beta - 1)^2}$	Range: (-∞, 1], unit less, larger is better. A diagonal decomposition of the NSE to separate correlation, bias and variability (Gupta et al., 2009).
Pearson correlation	$r_p = \frac{\sum_{i=1}^{N}(O_i - \bar{O})(P_i - \bar{P})}{\sqrt{\sum_{i=1}^{N}(O_i - \bar{O})^2}\sqrt{\sum_{i=1}^{N}(P_i - \bar{P})^2}}$	Range: [-1, 1], unit less, indicating the strength and direction of a linear relationship between two predicted and observed values. Close to 1 and -1 represent stronger positive and negative relationship, respectively. For predictive modelling, close to 1 is better.
Standard deviation ratio	$\alpha = \frac{Std(P)}{Std(O)}$	Range: (-∞, ∞), unit less, close to 1 is better.
Mean ratio	$\beta = \frac{\bar{P}}{\bar{O}}$	Range: (-∞, ∞), unit less, close to 1 is better.
Spearman correlation	$r_s = 1 - \frac{6\sum d_i^2}{N(N^2 - 1)}$	Range: [-1, 1], unit less, indicating the strength and direction of the monotonic relationship between two predicted and observed values. Close to 1 and -1 represent stronger positive and negative relationship, respectively. For predictive modelling, close to 1 is better.

N: Number of samples; O: Observed values; $\bar{O}$: Mean observed values; $Std(O)$: Standard deviation of observed values; P: Predicted values; $\bar{P}$: Mean predicted values; $Std(P)$: Standard deviation of predicted values; d_i^2: Difference between the ranks of each pair of observations

Code and data availability

Polished water table depth timeseries are available from Fang et al. (2025). Groundwater level simulations from DECIPHeR-GW are available from Zheng (2024). The meteorological forcings are available from CHESS (Robinson et al., 2023a; Robinson et al., 2023b). The surface elevation was extracted from the 10m LIDAR Composite Digital Terrain Model (DTM) produced by the Environment Agency in 2022 (https://www.data.gov.uk/dataset/7f31af0f-bc98-4761-b4b4-147bfb986648/lidar-composite-digital-terrain-model-dtm-10m). Height above nearest drainage (HAND) values were extracted from the global HAND dataset (Yamazaki et al., 2019). The soil attributes are available from the European Soil Database Derived data (https://esdac.jrc.ec.europa.eu/content/european-soil-database-derived-data). Spatial estimates of

https://doi.org/10.5194/egusphere-2026-3267
Preprint. Discussion started: 19 June 2026

transmissivity and specific yield across England and Wales were taken from Rahman et al. (2023). Population density and irrigation density were taken from Fang et al. (2026). Mean groundwater abstraction for the period 1999-2014 was taken from Rameshwaran et al. (2025).

The Python code developed in this study and the underlying data for model training and evaluation are available at https://github.com/QidongFang1203/LSTM_groundwater_modelling.

Supplement link

The link to the supplement will be included by Copernicus, if applicable.

Author contributions

QF: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualisation, Writing (original draft preparation). MR: Supervision, Conceptualization, Writing (review and editing). TW: Conceptualization, Writing (review and editing). FP: Supervision (lead), Conceptualization, Methodology, Writing (review and editing).

Competing interests

The authors declare that they have no conflict of interest.

Disclaimer

Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Acknowledgements

We greatly appreciate the discussions with Larisa Tarasova about the stochasticity in the LSTM network initialisation and optimisation, Xian Wang about LSTM network construction and optimisation, and Yanchen Zheng about the benchmarking with DECIPHeR-GW. We used ChatGPT and Gemini for language polishing of the part of this manuscript.

Financial support

Qidong Fang is funded by the China Scholarship Council (CSC) from the Ministry of Education of P.R. China. Thorsten Wagener acknowledges support from the Alexander von Humboldt Foundation in the framework of the Alexander von Humboldt Professorship endowed by the German Federal Ministry of Education and Research (BMBF). Francesca Pianosi is partially funded by the UK Engineering and Physical Sciences Research Council (grant EP/Y036999/1).

https://doi.org/10.5194/egusphere-2026-3267
Preprint. Discussion started: 19 June 2026

References

Allen, D. J., Brewerton, L. J., Coleby, L. M., Gibbs, B. R., Lewis, M. A., MacDonald, A. M., Wagstaff, S. J., and Williams, A. T.: The physical properties of major aquifers in England and Wales, British Geological Survey, Technical Report WD/97/34, 312 pp., http://nora.nerc.ac.uk/13137/, 1997.

Baste, S., Klotz, D., Acuña Espinoza, E., Bardossy, A., and Loritz, R.: Unveiling the limits of deep learning models in hydrological extrapolation tasks, Hydrol. Earth Syst. Sci., 29, 5871–5891, https://doi.org/10.5194/hess-29-5871-2025, 2025.

Baulon, L., Massei, N., Allier, D., Fournier, M., and Bessiere, H.: Influence of low-frequency variability on high and low groundwater levels: example of aquifers in the Paris Basin, Hydrol. Earth Syst. Sci., 26, 2829–2854, https://doi.org/10.5194/hess-26-2829-2022, 2022.

BGS (British Geological Survey): Groundwater resources in the UK, https://www.bgs.ac.uk/geology-projects/groundwater-research/groundwater-resources-in-the-uk/ (last access: 22 May 2026), 2024.

Bianchi, M., Scheidegger, J., Hughes, A., Jackson, C., Lee, J., Lewis, M., Mansour, M., Newell, A., O'Dochartaigh, B., Patton, A., and Dadson, S.: Simulation of national-scale groundwater dynamics in geologically complex aquifer systems: an example from Great Britain, Hydrol. Sci. J., 69, 572–591, https://doi.org/10.1080/02626667.2024.2320847, 2024.

Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J.: Classification and regression trees, 1st edn., Chapman and Hall/CRC, https://doi.org/10.1201/9781315139470, 1984.

Breiman, L.: Random forests, Mach. Learn., 45, 5–32, https://doi.org/10.1023/A:1010933404324, 2001.

Chidepudi, S. K. R., Massei, N., Jardani, A., Dieppois, B., Henriot, A., and Fournier, M.: Training deep learning models with a multi-station approach and static aquifer attributes for groundwater level simulation: what is the best way to leverage regionalised information?, Hydrol. Earth Syst. Sci., 29, 841–861, https://doi.org/10.5194/hess-29-841-2025, 2025.

Collenteur, R. A., Vonk, M. A., and Haaf, E.: Quantification and Analysis of Hydrograph Behavior Using Groundwater Signatures, Groundwater, 63, 779–789, https://doi.org/10.1111/gwat.13486, 2025.

Condon, L. E., Kollet, S., Bierkens, M. F. P., Fogg, G. E., Maxwell, R. M., Hill, M. C., Fransen, H.-J. H., Verhoef, A., Van Loon, A. F., Sulis, M., and Abesser, C.: Global groundwater modeling and monitoring: opportunities and challenges, Water Resour. Res., 57, e2020WR029500, https://doi.org/10.1029/2020WR029500, 2021.

Coxon, G., Addor, N., Bloomfield, J. P., Freer, J., Fry, M., Hannaford, J., Howden, N. J. K., Lane, R., Lewis, M., Robinson, E. L., Wagener, T., and Woods, R.: CAMELS-GB: hydrometeorological time series and landscape attributes for 671 catchments in Great Britain, Earth Syst. Sci. Data, 12, 2459–2483, https://doi.org/10.5194/essd-12-2459-2020, 2020.

Coxon, G., Freer, J., Lane, R., Dunne, T., Knoben, W. J. M., Howden, N. J. K., Quinn, N., Wagener, T., and Woods, R.: DECIPHeR v1: dynamic fluxEs and ConnectIvity for Predictions of HydRology, Geosci. Model Dev., 12, 2285–2306, https://doi.org/10.5194/gmd-12-2285-2019, 2019.

de Graaf, I. E. M., Sutanudjaja, E. H., Van Beek, L. P. H., and Bierkens, M. F. P.: A high-resolution global-scale groundwater model, Hydrol. Earth Syst. Sci., 19, 823–837, https://doi.org/10.5194/hess-19-823-2015, 2015.

de Graaf, I. E. M., van Beek, R. L. P. H., Gleeson, T., Moosdorf, N., Schmitz, O., Sutanudjaja, E. H., and Bierkens, M. F. P.: A global-scale two-layer transient groundwater model: development and application to groundwater depletion, Adv. Water Resour., 102, 53–67, https://doi.org/10.1016/j.advwatres.2017.01.011, 2017.

do Nascimento, T. V. M., Rudlang, J., Gnann, S., Seibert, J., Hrachowitz, M., and Fenicia, F.: Assessing the impact of geological map detail on process-based and data-driven hydrological models, Water Resour. Res., 62, e2025WR042375, https://doi.org/10.1029/2025WR042375, 2026.

Environment Agency: Hydrology Data Explorer, https://environment.data.gov.uk/hydrology/doc/reference (last access: June 2024), 2023.

Fan, Y., Li, H., and Miguez-Macho, G.: Global Patterns of Groundwater Table Depth, Science, 339, 940–943, https://doi.org/10.1126/science.1229881, 2013.

https://doi.org/10.5194/egusphere-2026-3267
Preprint. Discussion started: 19 June 2026

Fang, Q., Rahman, M., Wagener, T., Bloomfield, J. P., and Pianosi, F.: Widespread human footprints on groundwater trends across England, Environmental Research: Water, 2, 021001, https://doi.org/10.1088/3033-4942/ae61e9, 2026.

Fang, Q., Rahman, M., Wagener, T., Bloomfield, J., and Pianosi, F.: Groundwater Depth Time Series for 2902 wells across England, Zenodo [data set], https://doi.org/10.5281/zenodo.15584814, 2025.

Fisher, A., Rudin, C., and Dominici, F.: All models are wrong, but many are useful: learning a variable's importance by studying an entire class of prediction models simultaneously, arXiv [preprint], https://doi.org/10.48550/arXiv.1801.01489, 2019.

Freeze, R. A. and Cherry, J. A.: Groundwater, Prentice-Hall, Englewood Cliffs, NJ, ISBN 0-13-365312-9, 1979.

Fu, G., Crosbie, R. S., Barron, O., Charles, S. P., Dawes, W., Shi, X., Van Niel, T., and Li, C.: Attributing variations of temporal and spatial groundwater recharge: A statistical analysis of climatic and non-climatic factors, Journal of Hydrology, 568, 816–834, https://doi.org/10.1016/j.jhydrol.2018.11.022, 2019.

Gleeson, T., Wagener, T., Döll, P., Zipper, S. C., West, C., Wada, Y., Taylor, R., Scanlon, B., Rosolem, R., Rahman, S., Oshinlaja, N., Maxwell, R., Lo, M.-H., Kim, H., Hill, M., Hartmann, A., Fogg, G., Famiglietti, J. S., Ducharne, A., de Graaf, I., Cuthbert, M., Condon, L., Bresciani, E., and Bierkens, M. F. P.: GMD perspective: the quest to improve the evaluation of groundwater representation in continental- to global-scale models, Geosci. Model Dev., 14, 7545–7571, https://doi.org/10.5194/gmd-14-7545-2021, 2021.

Gnann, S., Baldwin, J. W., Cuthbert, M. O., Gleeson, T., Schwanghart, W., and Wagener, T.: The influence of topography on the global terrestrial water cycle, Rev. Geophys., 63, e2023RG000810, https://doi.org/10.1029/2023RG000810, 2025.

Goldstein, A., Kapelner, A., Bleich, J., and Pitkin, E.: Peeking inside the black box: visualizing statistical learning with plots of individual conditional expectation, J. Comput. Graph. Stat., 24, 44–65, https://doi.org/10.1080/10618600.2014.907095, 2015.

Green, T. R., Taniguchi, M., Kooi, H., Gurdak, J. J., Allen, D. M., Hiscock, K. M., Treidel, H., and Aureli, A.: Beneath the surface of global change: impacts of climate change on groundwater, J. Hydrol., 405, 532–560, https://doi.org/10.1016/j.jhydrol.2011.05.002, 2011.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling, J. Hydrol., 377, 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Heudorfer, B., Gupta, H. V., and Loritz, R.: Are deep learning models in hydrology entity aware?, Geophys. Res. Lett., 52, e2024GL113036, https://doi.org/10.1029/2024GL113036, 2025.

Heudorfer, B., Haaf, E., Stahl, K., and Barthel, R.: Index-based characterization and quantification of groundwater dynamics, Water Resour. Res., 55, 5575–5592, https://doi.org/10.1029/2018WR024418, 2019.

Heudorfer, B., Liesch, T., and Broda, S.: On the challenges of global entity-aware deep learning models for groundwater level prediction, Hydrol. Earth Syst. Sci., 28, 525–543, https://doi.org/10.5194/hess-28-525-2024, 2024.

Hochreiter, S. and Schmidhuber, J.: Long short-term memory, Neural Comput., 9, 1735–1780, https://doi.org/10.1162/neco.1997.9.8.1735, 1997.

Janssen, J., Tootchi, A., and Ameli, A. A.: Tackling water table depth modeling via machine learning: from proxy observations to verifiability, Adv. Water Resour., 201, 104955, https://doi.org/10.1016/j.advwatres.2025.104955, 2025.

Jasechko, S., Seybold, H., Perrone, D., Fan, Y., Shamsudduha, M., Taylor, R. G., Fallatah, O., and Kirchner, J. W.: Rapid groundwater decline and some cases of recovery in aquifers globally, Nature, 625, 715–721, https://doi.org/10.1038/s41586-023-06879-8, 2024.

Jiang, S., Sweet, L., Blougouras, G., Brenning, A., Li, W., Reichstein, M., Denzler, J., Shangguan, W., Yu, G., Huang, F., and Zscheischler, J.: How interpretable machine learning can benefit process understanding in the geosciences, Earth's Future, 12, e2024EF004540, https://doi.org/10.1029/2024EF004540, 2024.

https://doi.org/10.5194/egusphere-2026-3267
Preprint. Discussion started: 19 June 2026

Jones, H. K., Morris, B. L., Cheney, C. S., Brewerton, L. J., Merrin, P. D., Lewis, M. A., MacDonald, A. M., Coleby, L. M., Talbot, J. C., McKenzie, A. A., Bird, M. J., Cunningham, J., and Robinson, V. K.: The physical properties of minor aquifers in England and Wales, British Geological Survey, Technical Report WD/00/4, 234 pp., 2000.

Jung, H., Saynisch-Wagner, J., and Schulz, S.: Can eXplainable AI offer a new perspective for groundwater recharge estimation? Global-scale modeling using neural network, Water Resour. Res., 60, e2023WR036360, https://doi.org/10.1029/2023WR036360, 2024.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M.: Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks, Hydrol. Earth Syst. Sci., 22, 6005–6022, https://doi.org/10.5194/hess-22-6005-2018, 2018.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., and Nearing, G. S.: Toward improved predictions in ungauged basins: exploiting the power of machine learning, Water Resour. Res., 55, 11344–11354, https://doi.org/10.1029/2019WR026065, 2019b.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, Hydrol. Earth Syst. Sci., 23, 5089–5110, https://doi.org/10.5194/hess-23-5089-2019, 2019a.

Li, X., Khandelwal, A., Jia, X., Cutler, K., Ghosh, R., Renganathan, A., Xu, S., Tayal, K., Nieber, J., Duffy, C., Steinbach, M., and Kumar, V.: Regionalization in a global hydrologic deep learning model: from physical descriptors to random vectors, Water Resour. Res., 58, e2021WR031794, https://doi.org/10.1029/2021WR031794, 2022.

Liu, P. W., Famiglietti, J. S., Purdy, A. J., Adams, K. H., McEvoy, A. L., Reager, J. T., Bindlish, R., Wiese, D. N., David, C. H., and Rodell, M.: Groundwater depletion in California's Central Valley accelerates during megadrought, Nat. Commun., 13, 7825, https://doi.org/10.1038/s41467-022-35582-x, 2022.

Lundberg, S. and Lee, S.-I.: A unified approach to interpreting model predictions, arXiv [preprint], https://doi.org/10.48550/arXiv.1705.07874, 2017.

MacAllister, D. J., Krishan, G., Basharat, M., Cuba, D., and MacDonald, A. M.: A century of groundwater accumulation in Pakistan and northwest India, Nat. Geosci., 15, 390–396, https://doi.org/10.1038/s41561-022-00926-1, 2022.

Mackay, J. D., Jackson, C. R., and Wang, L.: A lumped conceptual model to simulate groundwater level time-series, Environ. Modell. Softw., 61, 229–245, https://doi.org/10.1016/j.envsoft.2014.06.003, 2014.

Margat, J. and van der Gun, J.: Groundwater around the world: a geographic synopsis, 1st edn., CRC Press, https://doi.org/10.1201/b13977, 2013.

Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C. A., Casalicchio, G., Grosse-Wentrup, M., and Bischl, B.: General pitfalls of model-agnostic interpretation methods for machine learning models, in: xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers, edited by: Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K.-R., and Samek, W., Springer International Publishing, 39–68, https://doi.org/10.1007/978-3-031-04083-2_4, 2022.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I: a discussion of principles, J. Hydrol., 10, 282–290, https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

Nobre, A. D., Cuartas, L. A., Hodnett, M., Rennó, C. D., Rodrigues, G., Silveira, A., Waterloo, M., and Saleska, S.: Height above the nearest drainage: a hydrologically relevant new terrain model, J. Hydrol., 404, 13–29, https://doi.org/10.1016/j.jhydrol.2011.03.051, 2011.

Nolte, A., Heudorfer, B., Bender, S., and Hartmann, J.: Multi-site deep learning for groundwater level prediction across global datasets: toward scalable applications under data scarcity, J. Hydroinform., jh2025095, https://doi.org/10.2166/hydro.2025.095, 2025.

Rahman, M., Pianosi, F., and Woods, R.: Simulating spatial variability of groundwater table in England and Wales, Hydrol. Process., 37, e14849, https://doi.org/10.1002/hyp.14849, 2023.

https://doi.org/10.5194/egusphere-2026-3267
Preprint. Discussion started: 19 June 2026

Rajaee, T., Ebrahimi, H., and Nourani, V.: A review of the artificial intelligence methods in groundwater level modeling, J. Hydrol., 572, 336–351, https://doi.org/10.1016/j.jhydrol.2018.12.037, 2019.

Rameshwaran, P., Bell, V. A., Davies, H. N., Sadler, P., Beverton, A., Thornton, R., and Rhodes-Smith, M.: Gridded actual groundwater, surface water and tidal water abstraction, discharge and Hands-off Flow datasets for England (1999 to 2014), NERC EDS Centre for Environmental Data Analysis [data set], https://doi.org/10.5285/18886f95ba84447f997efac96df456ad, 2025.

Reinecke, R., Foglia, L., Mehl, S., Trautmann, T., Cáceres, D., and Döll, P.: Challenges in developing a global gradient-based groundwater model (G3M v1.0) for the integration into a global hydrological model, Geosci. Model Dev., 12, 2401–2418, https://doi.org/10.5194/gmd-12-2401-2019, 2019.

Ribeiro, M. T., Singh, S., and Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144, https://doi.org/10.1145/2939672.2939778, 2016.

Robinson, E. L., Blyth, E. M., Clark, D. B., Comyn-Platt, E., Rudd, A. C., and Wiggins, M.: Climate hydrology and ecology research support system meteorology dataset for Great Britain (1961–2019) [CHESS-met], NERC EDS Environmental Information Data Centre [data set], https://doi.org/10.5285/835a50df-e74f-4bfb-b593-804fd61d5eab, 2023a.

Robinson, E. L., Blyth, E. M., Clark, D. B., Comyn-Platt, E., Rudd, A. C., and Wiggins, M.: Climate hydrology and ecology research support system potential evapotranspiration dataset for Great Britain (1961–2019) [CHESS-PE], NERC EDS Environmental Information Data Centre [data set], https://doi.org/10.5285/8651771d-aa6d-4d0f-8bcd-b3be1f733852, 2023b.

Rodell, M., Famiglietti, J. S., Wiese, D. N., Reager, J. T., Beaudoing, H. K., Landerer, F. W., and Lo, M. H.: Emerging trends in global freshwater availability, Nature, 557, 651–659, https://doi.org/10.1038/s41586-018-0123-1, 2018.

Rodell, M., Velicogna, I., and Famiglietti, J. S.: Satellite-based estimates of groundwater depletion in India, Nature, 460, 999–1002, https://doi.org/10.1038/nature08238, 2009.

Rushton, K. R.: Groundwater hydrology: conceptual and computational models, Wiley, 2004.

Siebert, S., Burke, J., Faures, J. M., Frenken, K., Hoogeveen, J., Döll, P., and Portmann, F. T.: Groundwater use for irrigation: a global inventory, Hydrol. Earth Syst. Sci., 14, 1863–1880, https://doi.org/10.5194/hess-14-1863-2010, 2010.

Slater, L., Blougouras, G., Deng, L., Deng, Q., Ford, E., Hoek van Dijke, A., Huang, F., Jiang, S., Liu, Y., Moulds, S., Schepen, A., Yin, J., and Zhang, B.: Challenges and opportunities of ML and explainable AI in large-sample hydrology, Philos. Trans. R. Soc. A, 383, 20240287, https://doi.org/10.1098/rsta.2024.0287, 2025.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res., 15, 1929–1958, 2014.

Tao, H., Hameed, M. M., Marhoon, H. A., Zounemat-Kermani, M., Heddam, S., Kim, S., Sulaiman, S. O., Tan, M. L., Sa'adi, Z., Mehr, A. D., Allawi, M. F., Abba, S. I., Zain, J. M., Falah, M. W., Jamei, M., Bokde, N. D., Bayatvarkeshi, M., Al-Mukhtar, M., Bhagat, S. K., Tiyasha, T., Khedher, K. M., Al-Ansari, N., Shahid, S., and Yaseen, Z. M.: Groundwater level prediction using machine learning models: a comprehensive review, Neurocomputing, 489, 271–308, https://doi.org/10.1016/j.neucom.2022.03.014, 2022.

Tarasova, L., Gnann, S., Yang, S., Hartmann, A., and Wagener, T.: Catchment characterization: Current descriptors, knowledge gaps and future opportunities, Earth-Science Reviews, 252, 104739, https://doi.org/10.1016/j.earscirev.2024.104739, 2024.

Williams, G. P.: Friends don't let friends use Nash-Sutcliffe Efficiency (NSE) or KGE for hydrologic model accuracy evaluation: a rant with data and suggestions for better practice, Environ. Modell. Softw., 194, 106665, https://doi.org/10.1016/j.envsoft.2025.106665, 2025.

Wunsch, A., Liesch, T., and Broda, S.: Deep learning shows declining groundwater levels in Germany until 2100 due to climate change, Nat. Commun., 13, 1221, https://doi.org/10.1038/s41467-022-28770-2, 2022.

https://doi.org/10.5194/egusphere-2026-3267
Preprint. Discussion started: 19 June 2026

Yamazaki, D., Ikeshima, D., Sosa, J., Bates, P. D., Allen, G. H., and Pavelsky, T. M.: MERIT Hydro: a high-resolution global hydrography map based on latest topography dataset, Water Resour. Res., 55, 5053–5073, https://doi.org/10.1029/2019WR024873, 2019.

Yang, C., Tijerina-Kreuzer, D. T., Tran, H. V., Condon, L. E., and Maxwell, R. M.: A high-resolution, 3D groundwater-surface water simulation of the contiguous US: advances in the integrated ParFlow CONUS 2.0 modeling platform, J. Hydrol., 626, 130294, https://doi.org/10.1016/j.jhydrol.2023.130294, 2023.

Yeh, H.-F., Cheng, Y.-S., Lin, H.-I., and Lee, C.-H.: Mapping groundwater recharge potential zone using a GIS approach in Hualian River, Taiwan, Sustainable Environment Research, 26, 33–43, https://doi.org/10.1016/j.serj.2015.09.005, 2016.

Zheng, Y., Coxon, G., Rahman, M., Woods, R., Salwey, S., Rong, Y., and Wendt, D. E.: DECIPHeR-GW v1: a coupled hydrological model with improved representation of surface–groundwater interactions, Geosci. Model Dev., 18, 4247–4271, https://doi.org/10.5194/gmd-18-4247-2025, 2025.

Zheng, Y.: DECIPHeR-GW v1: a coupled hydrological model with improved representation of surface-groundwater interactions, University of Bristol [data set], https://doi.org/10.5523/bris.wt0r1ec81zti2tww4p64fsqr3, 2024.