Journal of
Hydroinformatics

Journal of Hydroinformatics Vol 00 No 0, 1 doi: 10.2166/hydro.2025.095

Multi-site deep learning for groundwater level prediction across global datasets: toward scalable applications under data scarcity

Annika Nolte[a,b,*], Benedikt Heudorfer[c], Steffen Bender[a] and Jens Hartmann[b]

[a] Climate Service Center Germany (GERICS), Helmholtz-Zentrum Hereon, Hamburg, Germany
[b] Institute for Geology, Universität Hamburg, Bundesstraße 55, Hamburg 20146, Germany
[c] Karlsruhe Institute of Technology (KIT), Institute of Meteorology and Climate Research – Atmospheric Trace Gases and Remote Sensing, Karlsruhe, Germany
*Corresponding author. E-mail: annika.nolte@posteo.de

AN, 0000-0001-9562-0728

ABSTRACT

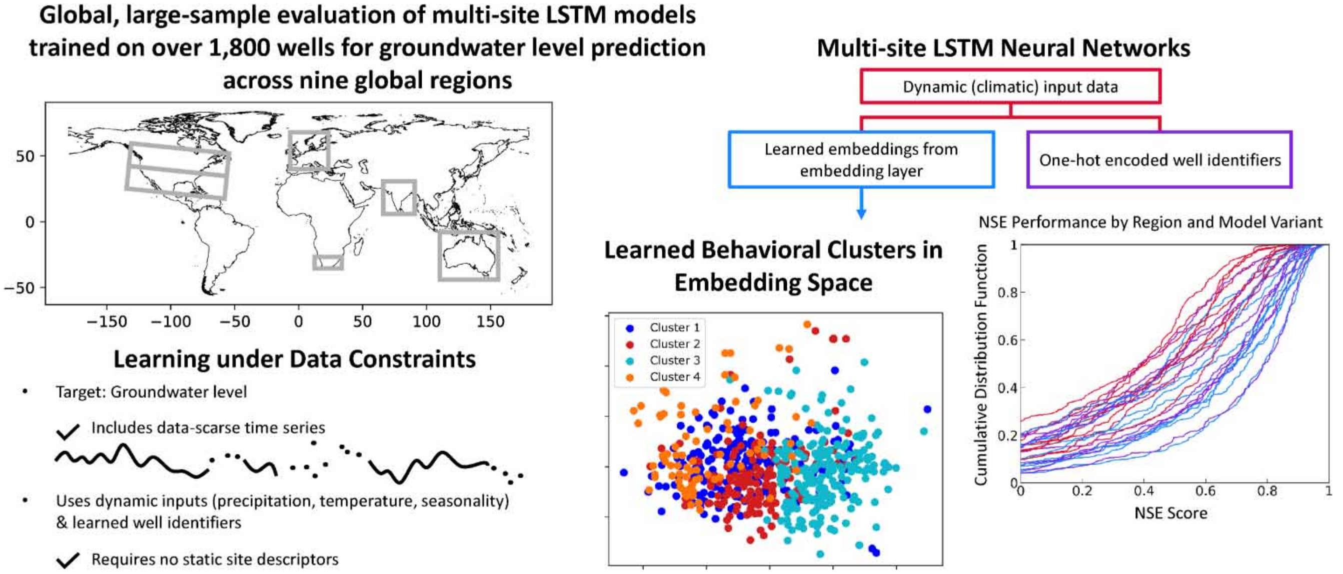
Deep learning (DL), and especially long short-term memory (LSTM) networks, have shown strong potential for groundwater level (GWL) prediction, but remain limited across sites and at scale due to data scarcity and reliance on sparse, inconsistent, or physically ambiguous static site descriptors. This study advances prediction in three key ways. First, we present the first global, large-sample evaluation of multi-site LSTM models trained on over 1,800 wells across nine global regions and evaluate their performance under realistic data constraints. Second, we show that trainable site embeddings enable the model to learn site-specific behavior directly from time series without relying on externally defined site descriptors. Third, we analyze the embedding space, revealing emergent spatial and functional patterns that reflect hydrogeological structure. Embedding-based models achieve strong predictive performance in data-rich regions (median Nash–Sutcliffe efficiency (NSE) $>$ 0.7) and remain robust across a combined global dataset. Our findings further challenge the assumption that larger datasets naturally improve predictions, especially in data-sparse regions. These findings describe both the potential and current limitations of scalable, entity-aware LSTM models for multi-site GWL prediction under data scarcity, paving the way for truly geographically global DL in hydrology.

Key words: data scarcity, deep learning, embeddings, entity-aware modeling, global, groundwater level prediction

HIGHLIGHTS

- Multi-site long short-term memory models were trained on 1,800 wells across nine global regions for groundwater level prediction.
- Embedding-based models outperformed others and achieved NSE $>$ 0.7 in data-rich regions.
- Global models matched regional performance, showing strong scalability.
- Site embeddings captured meaningful hydrogeological patterns without static input.
- This study shows promise for deep learning under data scarcity and diverse conditions.

GRAPHICAL ABSTRACT


Global, large-sample evaluation of multi-site LSTM models trained on over 1,800 wells for groundwater level prediction across nine global regions
50
0
−50
−150
−100
−50
0
50
100
150
Learning under Data Constraints
• Target: Groundwater level
✓ Includes data-scarse time series
• Uses dynamic inputs (precipitation, temperature, seasonality) & learned well identifiers
✓ Requires no static site descriptors
Multi-site LSTM Neural Networks
Dynamic (climatic) input data
Learned embeddings from embedding layer
One-hot encoded well identifiers
Learned Behavioral Clusters in Embedding Space
Cluster 1
Cluster 2
Cluster 3
Cluster 4
NSE Performance by Region and Model Variant
Cumulative Distribution Function
NSE Score
0
0.2
0.4
0.6
0.8
1

1. INTRODUCTION

Groundwater supplies water for half of the world's domestic use and 38% of the world's irrigated land (United Nations 2022). As part of the hydrological cycle, groundwater influences climate regulation and supports biodiversity and ecosystem services (Gleeson *et al.* 2020b). It requires observational data and modeling techniques to gain an understanding of past and future trends in groundwater recharge and availability (Sanford 2002; Scanlon *et al.* 2002). Currently, *in-situ* monitoring provides the most accurate insights into groundwater availability, with groundwater level (GWL) data being the most commonly available parameter, enabling groundwater modeling and trend estimations. However, GWL data often remains sparse in many places (Lall *et al.* 2020; United Nations 2022), and GWL prediction is further complicated by the inherent complexity of its dynamics and the numerous components and interactions within groundwater systems, driven by both climate and human activities (de Graaf *et al.* 2019; Gleeson *et al.* 2020a).

Traditional numerical groundwater model approaches approximate or simplify system complexity. Data scarcity, as well as the time-consuming data standardization and integration process, often limit these approaches. In addition, addressing the potential impacts of climate change on future groundwater resources generally requires high-quality, long-term, continuous GWL time series, which are rarely available in many regions worldwide. Therefore, GWLs are often analyzed with a limited number of representative wells for large regions (e.g., Bloomfield *et al.* 2019; Cuthbert *et al.* 2019b). This underscores the necessity of bridging the gap between the consideration of localized hydro(geo)logic processes based on local data and large-scale effects, a critical step for water management in practice (United Nations 2022). It calls for innovative approaches that can effectively scale predictions by integrating diverse datasets and capturing complex interactions across different spatial scales (Massei *et al.* 2020; Gleeson *et al.* 2021). In this regard, data-driven algorithms have significantly advanced, with deep learning (DL), a subset of machine learning (ML) characterized by its use of layered neural networks, emerging as a particularly promising tool for accurate GWL prediction with minimal *a priori* process understanding at well locations (Rajaee *et al.* 2019; Tao *et al.* 2022). Beyond groundwater, DL and other ML approaches have been widely applied in hydrology, including rainfall forecasting (Mekanik *et al.* 2013), streamflow and runoff prediction (Kratzert *et al.* 2019; Balacumaresan *et al.* 2024), water supply and demand modeling (Mohammadi *et al.* 2024), and water quality monitoring (Najah *et al.* 2013), demonstrating their broad potential for capturing nonlinear hydrological processes across domains (Shen & Lawson 2021).

Early DL efforts in surface and subsurface hydrology were primarily focused on training models for individual wells with carefully selected high-quality time series (Lee *et al.* 2019; Solgi *et al.* 2021; Wunsch *et al.* 2022; Chidepudi *et al.* 2023). More recently, attention has shifted toward multi-site DL architectures – specifically long short-term memory (LSTM) networks (Hochreiter & Schmidhuber 1997) – trained on large, spatially distributed datasets, enabling simultaneous simulation of hydrological behavior across multiple locations. In this context, neural networks are trained on time series from numerous spatially distributed sites, allowing them to learn shared patterns while retaining local specificity. A key methodological advancement enabling this transition has been the introduction of entity-aware modeling by Kratzert *et al.* (2018, 2019), which equips models with the ability to distinguish between individual sites, or 'entities', during training and prediction. This is typically achieved by providing the models with two types of input: dynamic input features such as meteorological forcing and static features relevant to describing the physical environment (e.g., hydrogeological and topographic attributes). These static features are intended to inform the model about inter-site (dis)similarities in hydrological functioning, thereby guiding the learning of inter-site similarities and improving predictive skill across sites (Kratzert *et al.* 2019; Heudorfer *et al.* 2024; Kunz *et al.* 2024).

In surface hydrology, standardized datasets like CAMELS (Addor *et al.* 2017) provide well-tested, spatially aligned static features that support this modeling paradigm for large spatial scales. In hydrogeology, however, no comparable global or continental-scale datasets exist. Groundwater systems are characterized by high spatial heterogeneity, limited observability, and often incomplete or uncertain information regarding aquifer properties, geology, or anthropogenic impacts. Furthermore, the derivation of static features frequently relies on proximity-based proxies (e.g., using nearby or regional land cover, soil, or topography to represent local hydrogeological properties) or aggregated statistics within spatial buffer zones around wells to capture environmental characteristics, which can introduce additional uncertainty due to spatial heterogeneity and limited representativity of these features for subsurface dynamics (Haaf *et al.* 2023; Heudorfer *et al.* 2024; Kunz *et al.* 2024). Recent studies have begun to question the added value of such static features in LSTM-based hydrological models. Evidence from both groundwater and surface water applications shows that time series-derived statistics, random site identifiers, or one-hot encoded site labels can yield predictive performance comparable to that of models using detailed environmental descriptors, implying that static features often function more as unique entity identifiers than as meaningful representations of hydrological or hydrogeological processes (Li *et al.* 2022; Heudorfer *et al.* 2024, 2025). Under this assumption, their absence may not substantially reduce model performance, raising the question of whether models can instead learn site-specific behavior and shared representations directly from time series.

In this study, we follow this assumption and investigate whether LSTM-based multi-site GWL prediction models can learn site-specific behavior and potentially shared representations directly from dynamic inputs and site identifiers, without relying on predefined static descriptors. Our approach is based on modeling and evaluating GWL time series from over 1,800 wells across five continents, encompassing a wide range of data quality, record lengths, and sampling frequencies. In contrast to prior studies using high-quality or pre-screened time series (e.g., Heudorfer *et al.* 2024; Chidepudi *et al.* 2025), we evaluate the performance of the model under realistic global conditions. This includes time series of varying completeness, noise levels, and hydrological contexts, which are conditions commonly encountered in practice but rarely modeled. In doing so, we address a critical question: whether multi-site LSTM models can offer accurate, scalable solutions for groundwater prediction under data scarcity conditions – both in terms of time series and static feature availability.

As part of our investigation, we implement embeddings as a form of self-learned entity identification, assigning each monitoring well a trainable numerical vector that is updated during model training. This enables the model to capture site-specific behavior and inter-site similarity without requiring explicit physical metadata – unlike one-hot encodings, which treat all sites as unrelated (e.g., Chidepudi *et al.* 2025). The approach is comparable to the process by which embeddings in natural language processing capture contextual relationships between words (Akbik *et al.* 2018; Wang *et al.* 2020). A similar mechanism was applied by Clark *et al.* (2022) in a multi-site DeepAR model for GWL prediction, where categorical embeddings were used to represent individual wells. However, the structure and meaning of the learned embeddings were not analyzed or interpreted by the authors. In contrast, our study introduces an explicit, trainable embedding layer and evaluates its utility for site differentiation and the internal representation of cross-site variability under data-scarce conditions. This assessment is inherently constrained by the same lack of independent physical descriptors that initially motivated the use of embeddings over predefined static features. Recent work in surface hydrology has demonstrated that learned internal representations in DL models can capture meaningful hydrological behavior. Lees *et al.* (2021) suggested that LSTM cell states can reflect soil moisture dynamics, while Bassi *et al.* (2024) used autoencoder-derived landscape fingerprints to explain streamflow variability. Botterill & McMillan (2023) developed convolutional encoders to derive low-dimensional hydrological 'signatures' from time series, offering an interpretable alternative to classical catchment descriptors.

The rest of this paper is organized as follows. Section 2 introduces the data and well locations. Section 3 outlines our methodology, employing LSTMs in an explorative approach, where we use this model architecture with either only dynamic input data or with additional site identifiers provided through one-hot encoding or trainable embeddings. Section 4 presents the results and discussion, focusing on both model performance and the explanation of embeddings to address our study objectives and their broader implications. Finally, Section 5 concludes the study by outlining its contributions to advancing large-scale environmental modeling using DL, particularly in the context of data scarcity.

2. STUDY AREA AND DATA

Precipitation and temperature data were used as dynamic input features and originated from the ERA5 dataset, the fifth generation of atmospheric reanalysis by the European Centre for Medium-Range Weather Forecasts (ECMWF) (Hersbach *et al.* 2020). Additionally, we used a third dynamic input feature that has proven valuable for GWL prediction in previous studies (Wunsch *et al.* 2021; Heudorfer *et al.* 2024). This feature involved fitting an annual sinusoidal curve to the temperature data, which potentially allows the model to capture seasonal patterns more effectively.

The modeled GWL time series originate from a global dataset previously compiled to investigate coastal GWL dynamics within 100 km of the shoreline (Nolte *et al.* 2024). The corresponding data sources are listed in Supplementary Table S1. For this study, we selected shallow wells with a maximum median water table depth of 20 m to focus on near-surface conditions. The CoastalDEM elevation dataset (Kulp & Strauss 2019) was used to convert groundwater elevation data to GWLs referenced to the ground surface (if applicable), ensuring a consistent reference point across the dataset. Wells in semi-confined or confined aquifers were excluded when such information was available, as these aquifers often show different hydraulic behaviors compared with unconfined aquifers. This makes their dynamic features less comparable. From this data pool, GWL time series were selected and preprocessed. We chose GWL data from the period 1979 to 2018, because this period aligns with the ERA5 climate data availability. The time series were aggregated to a weekly resolution and clustered with respect to geographic regions, creating nine datasets with over 200 GWL time series each.

The following selection criteria were applied, with specific values provided in Table 1: (1) the time series length in years (length); (2) the proportion of the data points in weekly aggregated time series that have GWL data available (availability); and (3) the maximum duration, measured in weeks, during which no GWL data is available (gap length). In some cases, the data selection includes only segments of the time series, i.e., not necessarily the complete time series. Time series selected using these criteria then underwent visual screening – a step still recommended to identify and minimize errors, noise, and anthropogenic effects inside the GWL data (Barthel *et al.* 2022; Retike *et al.* 2022). We found data points significantly outside the value range (outliers), sudden sharp level changes over specific periods, and patterns potentially indicative of well pumping or anthropogenic recharge events, guided by our own experience and illustrative examples shown in Retike *et al.* (2022). Additionally, we applied Density-Based Spatial Clustering of Applications with Noise (DBSCAN; Ester *et al.* 1996) to rescue a few GWL time series by grouping similar data points into clusters based on their spatial and temporal proximity while identifying and excluding outliers. Finally, we randomly selected 200 time series for each refined regional dataset to ensure equally large sample sizes for every region and enhance comparability. These 200 wells were drawn from initial regional pools ranging from approximately 280 to 480 wells each. Geographical delineations of focus regions are illustrated in Figure 1(a). Based on the nine focus regions (Figure 1(b)), the full, global dataset includes 1,800 GWL time series that vary in data quantity, including periods covered, data gaps, and overall sample sizes (Supplementary Figure S1). The focus regions also span a wide range of climatic conditions (Supplementary Figure S2). Rather than being perfectly physiographically

Table 1 | Criteria for selecting GWL time series per focus region (Figure 1(a)), including length, availability, and gap length; median GWL per focus region

Criteria	AU	DE	FR	IN	NA-N	NA-S	NL	Scand	ZA
1. Length (years)	9	15	6[a]	15	12	12	12[a]	15	4
2. Availability (%)	40	90	90	7	20	20	90	20	20
3. Gap length (weeks)	12	2	2	26	12	12	2	12	12
GWL median	−5.0	−4.5	−5.0	−4.6	−3.8	−3.2	−1.2	−2.9	−4.5

More information on the selection criteria is available in Section 2.

[a] Initially, about 30 GWL time series with a minimum length of 15 years were chosen. Afterward, the remaining GWL time series that met the length criteria given in the table were randomly selected from the dataset.

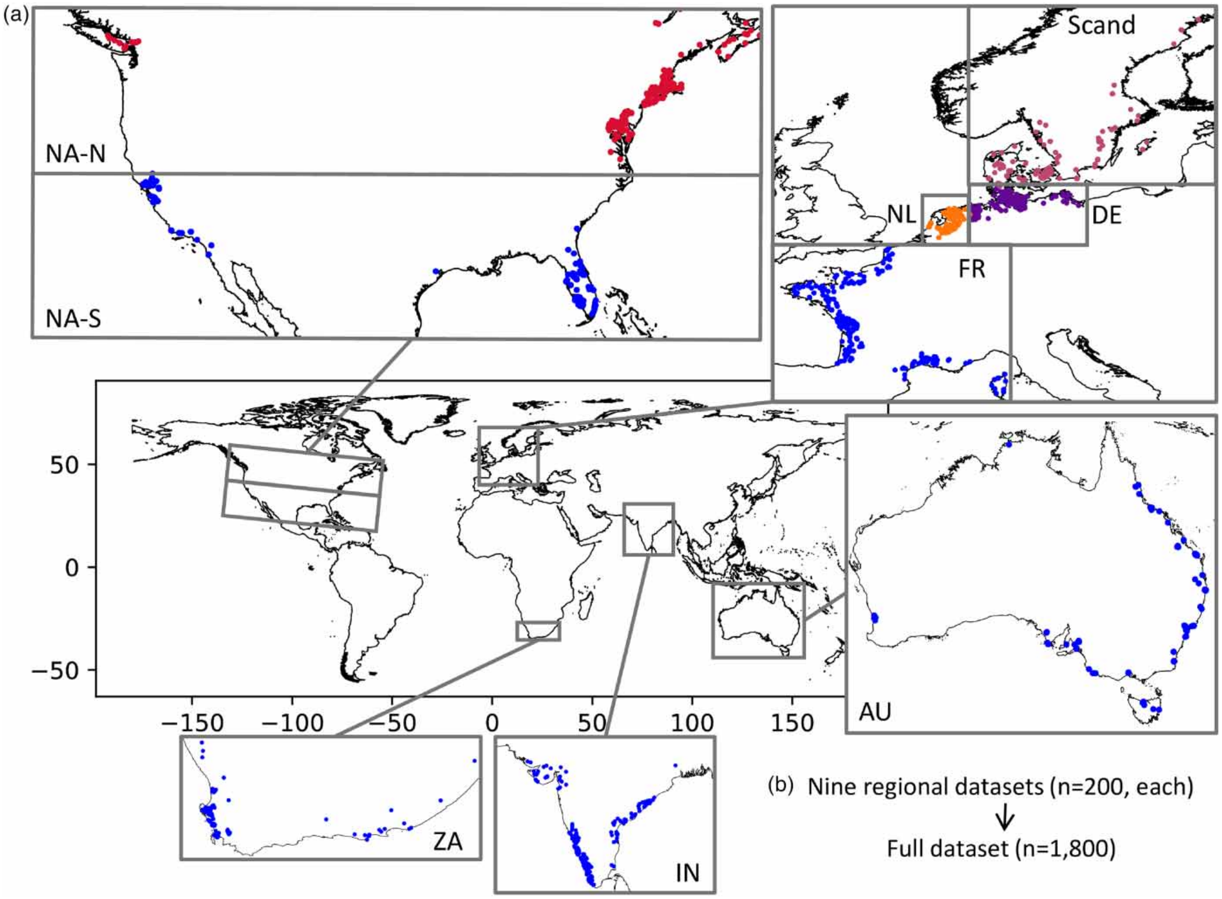


Figure 1 | (a) Maps of the focus regions highlighting the geographical areas from which regional datasets have been sourced for this study: Australia (AU); Germany (DE); France (FR); India (IN); North America – North (NA-N); North America – South (NA-S); Netherlands (NL); Scandinavia containing Denmark and Sweden (Scand); South Africa (ZA). Colored dots represent the well locations of selected GWL time series. (b) Overview of model application scale. Each of the nine regional datasets contains 200 wells, collectively forming the full, global dataset used in the global modeling experiments ($n = 1,800$).

representative or reflective of typical data quality and quantity in their respective regions, the datasets reflect the diverse and often inconsistent nature of global groundwater datasets.

For the modeling task, each time series dataset at the selected well locations (GWL, precipitation, temperature, and temperature seasonality) was standardized to have a mean of zero and a standard deviation of one, addressing discrepancies in scale and the impact of errors across different time series. This standardization, which was fitted on the training portion of the data (Section 3.2), ensures that no single series dominates the learning process.

3. METHODS

3.1. Scope and experimental design

This study focuses on model development for GWL prediction by evaluating the utility of LSTM-based models in data-scarce, large-scale settings, where both time series and static input features are limited. While benchmarking against process-based, hybrid, or alternative ML models remains an active area of research, previous studies have demonstrated that LSTMs often outperform (i.e., achieve higher predictive accuracy, such as improved NSE or lower RMSE when capturing observed groundwater or streamflow dynamics) or complement traditional approaches for hydrological time series prediction and are widely regarded as a state-of-the-art, well-established DL method in this context (Kratzert *et al.* 2019; Rajaee *et al.* 2019; Lees *et al.* 2021). Building on these findings, we assess LSTM performance under real-world data constraints and investigate the role of trainable embeddings for representing site-specific behavior.

We tested the prediction capabilities of LSTM models within a multi-site architecture in a series of experiments using two setups (Figure 2): a standard configuration (Section 3.2) and a modified setup with embeddings (Section 3.3). Both model setups were applied to each of the individual regional datasets as well as to the global dataset (Figure 1(b)). In the standard setup, we used either exclusively dynamic input data comprising precipitation and temperature (*nofeat* models), or static site information in the form of one-hot-encoded well identifiers (ohe features; *ohefeat* models), which were concatenated to the dynamic inputs at each time step. This allows the multi-site model to distinguish between different groundwater dynamics of individual wells. Similarly, the embedding-based setup (*embeddings* models) uses well identifiers to generate trainable vectors that serve as compact representations of individual wells. In this setup, the matrix from the static input has the dimension of the embeddings instead of being multiplied by the number of wells as in the *ohefeat* models. To explore how embeddings might enhance model interpretability and capture spatially meaningful patterns, we applied dimensionality reduction and clustering techniques (Section 3.3). We structured the evaluation of our modeling framework around three main analytical dimensions:

1. *Model architecture comparison* (Sections 4.1 and 4.2): We compared LSTM models with different site representations (no static input, one-hot encoding, and embeddings) across all regional datasets. This allowed us to assess the role of dynamic and site-specific information and evaluate the benefit of embedding-based representations.
2. *Training data scope, quantity, and cross-regional learning* (Sections 4.3 and 4.4): We evaluated the effects of data volume, diversity, and completeness by comparing regionally trained models among themselves and against a global model, assessing trade-offs between specialization and the integration of diverse datasets under varying data conditions.
3. *Learned spatial representations and internal differentiation across sites* (Section 4.5): We analyzed the embedding space to assess whether site-specific groundwater dynamics were internally encoded in a way that reflects spatial and functional similarity, enabling differentiation across wells without predefined static features.

3.2. Standard LSTM

We applied a standard LSTM architecture in which all input features enter the model through the same input layer (Hochreiter & Schmidhuber 1997). For more details on the model architecture, we refer to previous applications, such as those by Kratzert *et al.* (2019). The LSTM model was trained to make predictions at a weekly resolution.

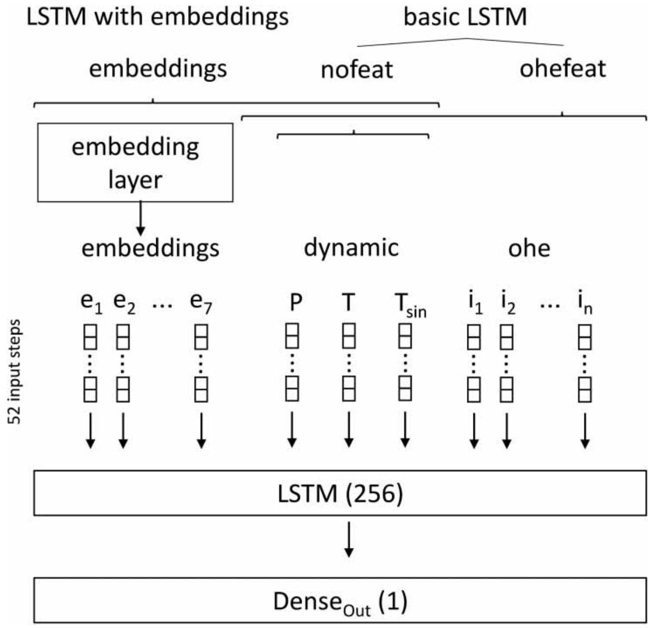


Figure 2 | LSTM network model structure for the three model variants: *nofeat* (three dynamic input features only), *ohefeat* (dynamic inputs combined with one-hot encoded well identifiers *i*, increasing the input dimensionality by *n*, the number of individual GWL time series), and *embeddings* (dynamic input features combined with seven learned well embeddings *e*).

In the multi-site architecture, data from multiple wells were combined: all time series were aligned chronologically, and missing values were retained as gaps. Each GWL time series was split into three set parts for training (70%), validation (15%), and testing (15%). We then pooled the sequences from all wells into three 'buckets' (train/stop/test). We conducted a preliminary investigation on the validation set to assess general performance and examine the effects of the newly implemented embeddings (Section 3.3).

Model performance was primarily assessed using the Nash–Sutcliffe efficiency (NSE), calculated on the independent test set to ensure unbiased evaluation. High NSE values (closer to 1) indicate strong agreement between model predictions and observed variance during the test period. In addition to NSE, we computed several complementary metrics: Kling–Gupta efficiency (KGE), coefficient of determination (R^2), bias, mean squared error (MSE), and root mean squared error (RMSE). However, NSE was used as the primary performance indicator throughout the analysis.

LSTM networks are designed to capture long-term dependencies in time series data through memory cells and gating mechanisms, making them well-suited for irregular temporal patterns. However, they do not inherently handle missing values, which must be explicitly addressed during training. We applied a masking technique that excluded missing values from error calculations, ensuring that training and evaluation focused only on available data points.

The MSE served as the loss function in the training process. During training, the Adam optimizer (Kingma & Ba 2014) was used to find the set of weights that minimizes this loss. The standardization of the GWL time series (see Section 2) prevented specific wells, particularly those with larger fluctuations or inherently larger values, from disproportionately influencing the loss. Standardization also aligned the scale of the MSE used in training with that of the NSE used in testing, ensuring that error metrics were consistent and comparable across wells in a multi-site setup (Heudorfer *et al.* 2024).

Hyperparameters control both the architecture and the learning procedure of the LSTM. Following common practice in large-sample hydrology, we used a single-layer LSTM with 256 hidden units, followed by dropout and a dense output layer (Figure 2; full settings in Supplementary Section C). This architecture was originally selected in Kratzert *et al.* (2019) via grid search with cross-validation on the CAMELS basins, has therefore been applied across diverse hydroclimatic settings, and is now a common baseline (e.g., Frame *et al.* 2022; Klotz *et al.* 2022; Loritz *et al.* 2024; Heudorfer *et al.* 2025). To keep model capacity appropriate for dataset size and aligned with prior applications of this architecture (typically mid-hundreds of basins/wells), we limited each regional dataset to 200 wells so that regional and global experiments were on a comparable scale. We did not perform additional hyperparameter optimization to maximize comparability across setups and datasets, and because we aimed to understand the capabilities and limitations of the LSTM models rather than to push absolute performance. In this context, further tuning typically brings incremental improvements rather than categorical differences when the model is already reasonably sized, i.e., not excessively large or small.

To assess the robustness of the model results, each of the three model variants (Figure 2) was trained and evaluated 10 times using different random seeds, resulting in distinct weight and bias initializations. While more advanced uncertainty quantification methods (e.g., Monte Carlo dropout) exist (Klotz *et al.* 2022), this approach follows standard practice and is sufficient for model robustness evaluation in similar hydrological DL studies (e.g., Kratzert *et al.* 2019; Chidepudi *et al.* 2025; Mangukiya & Sharma 2025). It allowed us to evaluate performance variability across training runs and ensured that findings were not driven by a single initialization.

3.3. LSTM with embeddings

The model architecture in the LSTM with the *embeddings* model setup is similar to that in the standard LSTM model setup described in Section 3.2, with one exception: the substitution of the one-hot encoding with an embedding layer (Figure 2). Like the *ohefeat* models, the encoding process of the *embeddings* models differentiates between individual wells in the dataset. However, well identifiers are converted into a dense, lower-dimensional space in the embedding layer. Specifically, we employed a seven-dimensional embedding space for representing well identifiers in our LSTM models. This dimensionality was chosen to balance computational efficiency and the capability to capture and distinguish well characteristics across the large dataset. By doing so, it greatly reduces the number of input dimensions compared with one-hot encoding.

For each well identifier, an embedding is retrieved from an embedding matrix E, where each row corresponds to a unique well identifier. This can be expressed as:

$$E_i = E[X_{\text{static}}[i]]$$

where E_i represents the embedding for the ith well identifier in the input vector X_{static}. While these embeddings do not encode wells uniquely in the way one-hot encoded vectors do, they may help the model capture shared behavior patterns among wells based on their time series.

The embedding layer, implemented using the Python library Keras, learns an embedding matrix during training to optimize the representations of the individual wells for the task. The embeddings are adjusted during the training process to minimize the prediction error by capturing similarities in patterns within the time series. Specifically, the embeddings are optimized to predict GWLs using combined static (embedding-derived) and dynamic (meteorological) inputs. After this transformation, both inputs are concatenated and fed into the LSTM layer, similar to the *ohefeat* model variant (Figure 2).

3.4. Exploring learned representations through embeddings

The embedding layer (Section 3.3) provides an opportunity to investigate how the model organizes information about individual wells based on their time series. LSTM models are often regarded as 'black boxes' due to their complex internal states and lack of direct interpretability. Although embeddings do not constitute explainability in the classical sense of feature attribution (e.g., SHAP or LIME), they allow for *post hoc* analysis of how the model internally differentiates between sites.

To investigate these learned representations, we analyzed the distribution of the embeddings values in relation to geographical location and GWL dynamics, which have previously been linked to environmental descriptors such as topography and lithology (Nolte *et al.* 2024). Both linear (principal component analysis, PCA) and nonlinear (t-distributed stochastic neighbor embedding, t-SNE; van der Maaten & Hinton 2008) dimensionality reduction techniques were used to project the seven-dimensional embedding space into two dimensions for visual exploration. The resulting projections were then visualized using color-coding to assess whether the embeddings captured meaningful structure. Two coloring strategies were employed: (a) based on geographical information, to reflect large-scale climatic influences and (b) based on clusters of GWL dynamics previously defined by Nolte *et al.* (2024). For (a), wells were categorized by geographical regions (continents and countries). For (b), wells were assigned to one of four clusters derived from statistical patterns in hydrographs (Heudorfer *et al.* 2019) representing different types of GWL behavior (e.g., stable seasonal cycles, rapid recharge responses, and interannual variability). These clusters were based on approximately 8,000 globally distributed groundwater hydrographs and included nearly all wells from the DE dataset and about 35% of the wells in the full dataset of this study.

Beyond visual comparisons, a quantitative method can help to assess the alignment between the learned embeddings and the prior GWL dynamics clusters. Specifically, we compared the four prior clusters with four newly generated clusters obtained by applying k-means clustering to the embeddings (using the same settings and number of principal components as in the prior study). Cluster cohesion and separation was evaluated using silhouette scores (Rousseeuw 1987), where higher values indicate more internally consistent clusters. In addition, the normalized mutual information (NMI; Strehl & Ghosh 2002) was used to compare both sets of clusters, i.e., to quantify the correspondence between the learned and reference clusters. In addition, we used DBSCAN (eps = 0.25, min_samples = 2) to identify inliers and outliers in the embedding space, aiming to detect wells that were isolated or underrepresented in the learned representation. The subsequent comparison of model performance metrics between these groups aimed to assess whether sparsely represented wells were associated with lower predictive skill.

4. RESULTS AND DISCUSSION

4.1. Model performance across regions

This section presents the results of LSTM models trained and tested on regional datasets only. Across the nine regions, a total of 270 model runs were conducted, covering three model variants and accounting for variability through multiple training initializations. The resulting model performances, summarized in Table 2 and Figure 3, show how predictive skill varies by region and input configuration. Spatial patterns in model accuracy across individual wells are illustrated in Supplementary Section D, showing regional and local variation within the datasets.

Models using embeddings and one-hot encoded site identifiers (*ohefeat*) consistently outperformed the variant without static input features (*nofeat*), as discussed in more detail in Section 4.2. The *embeddings* variant achieved the highest median performance in eight out of nine regions. In five regional datasets – AU, DE, FR, NA-N, and NL – these models achieved moderate to high performance, with median NSE scores ranging between 0.67 and 0.80. In contrast, moderate to low performance, with median NSE scores from 0.40 to 0.55, was derived for the datasets from IN, NA-S, Scand, and ZA. Regions with higher performance also exhibited lower variability in the models with site identifiers, as reflected by

Table 2 | Mean, 10th percentile (Q10), and 90th percentile (Q90) NSE scores across 10 random initializations, evaluated for the three model variants *nofeat*, *embeddings*, and *ohefeat* (Figure 2) across the nine focus regions shown in Figure 1(a)

NSE	Q10	Q50	Q90	Q10	Q50	Q90	Q10	Q50	Q90
	AU			**DE**			**FR**		
nofeat	0.425	0.495	0.544	0.399	0.462	0.529	0.549	0.594	0.626
embeddings	**0.670**	**0.701**	**0.739**	**0.708**	**0.729**	**0.746**	0.774	0.788	0.798
ohefeat	0.637	0.670	0.691	0.664	0.701	0.745	**0.781**	**0.795**	**0.806**
	IN			**NA-N**			**NA-S**		
nofeat	**0.489**	0.509	0.524	0.522	0.536	0.554	0.382	0.398	0.414
embeddings	0.488	**0.523**	**0.552**	**0.679**	**0.692**	**0.701**	**0.437**	**0.459**	**0.477**
ohefeat	0.448	0.466	0.489	0.674	0.682	0.690	0.381	0.404	0.421
	NL			**Scand**			**ZA**		
nofeat	0.424	0.488	0.562	0.362	0.385	0.415	0.369	0.398	0.440
embeddings	**0.674**	**0.688**	**0.708**	**0.569**	**0.585**	**0.609**	0.496	**0.599**	**0.667**
ohefeat	0.655	0.675	0.698	0.446	0.492	0.537	**0.509**	0.552	0.589

Additional evaluation metrics are available in Supplementary Table S2. Bold values indicate the highest NSE score among the three model variants (*nofeat*, *embeddings*, *ohefeat*) for each region and quantile.

smaller interquartile ranges (Q90–Q10). For instance, in FR, the *embeddings* and *ohefeat* models both achieved stable predictive skill, with NSE values ranging between 0.78 and 0.81. Additional evaluation metrics (Supplementary Table S2) support these findings.

4.2. Superior performance of entity-aware models

While dynamic input features, which capture temporal patterns and enable learning from historical variability, are most important for GWL prediction, the inclusion of direct site identifiers in the *ohefeat* and *embeddings* models further improved performance by 16–37% (Table 2; Figure 3). Except for the datasets from IN and NA-S, the *nofeat* models show notably lower performance compared with the model variants with site identification. Among these, the *embeddings* models outperformed the *ohefeat* models across all datasets except FR, with median NSE improvements of up to 0.1 in several datasets.

Our findings of enhanced predictive capabilities in models with direct site identifiers align with previous studies that demonstrated the success of these models in learning hydrological patterns across catchments in the United States (Kratzert *et al.* 2019) and GWL dynamics in Germany (Heudorfer *et al.* 2024; Kunz *et al.* 2024). These models, often referred to as entity-aware, can help overcome current limitations posed by uncertainties in descriptive environmental data. However, we observed only marginal performance improvements from including site identifiers in the IN dataset. In this case, the GWL data are particularly sparse but regularly distributed (Supplementary Figure S1), suggesting a potential upper limit to the ability of the model to discern patterns at the given temporal resolution and data quality (compare Section 4.4). Consequently, a minimum monthly data frequency appears necessary to support the effective use of site-specific static input.

Performance benefits with site identifiers are evident for most wells. The cumulative distribution functions (CDFs) in Figure 3(a) show a steeper ascent for the *embeddings* and *ohefeat* models compared with the *nofeat* model, indicating a higher proportion of wells achieving elevated NSE scores when site identification is included. Wells with medium to high baseline scores exhibit noticeable improvements in predictive accuracy. In contrast, wells with initially low scores show little to no performance gain, suggesting that modeling challenges in these cases are primarily driven by limited or compromised data. Similarly, a subset of wells with negative NSE values – indicating less accurate predictions than the average of the observed data – remain difficult to model across all model variants. This is consistent with previous multi-site LSTM studies in both surface and groundwater hydrology (Kratzert *et al.* 2019; Li *et al.* 2022; Heudorfer *et al.* 2024).

The *ohefeat* model requires substantially more computational resources due to increased input dimensionality, while the *embeddings* model provides a more resource-efficient alternative by reducing input space without compromising performance. This makes the *embeddings* approach particularly advantageous for large-sample datasets with extensive GWL time series and for large-scale applications, such as the global dataset used in this study (see Section 4.3).

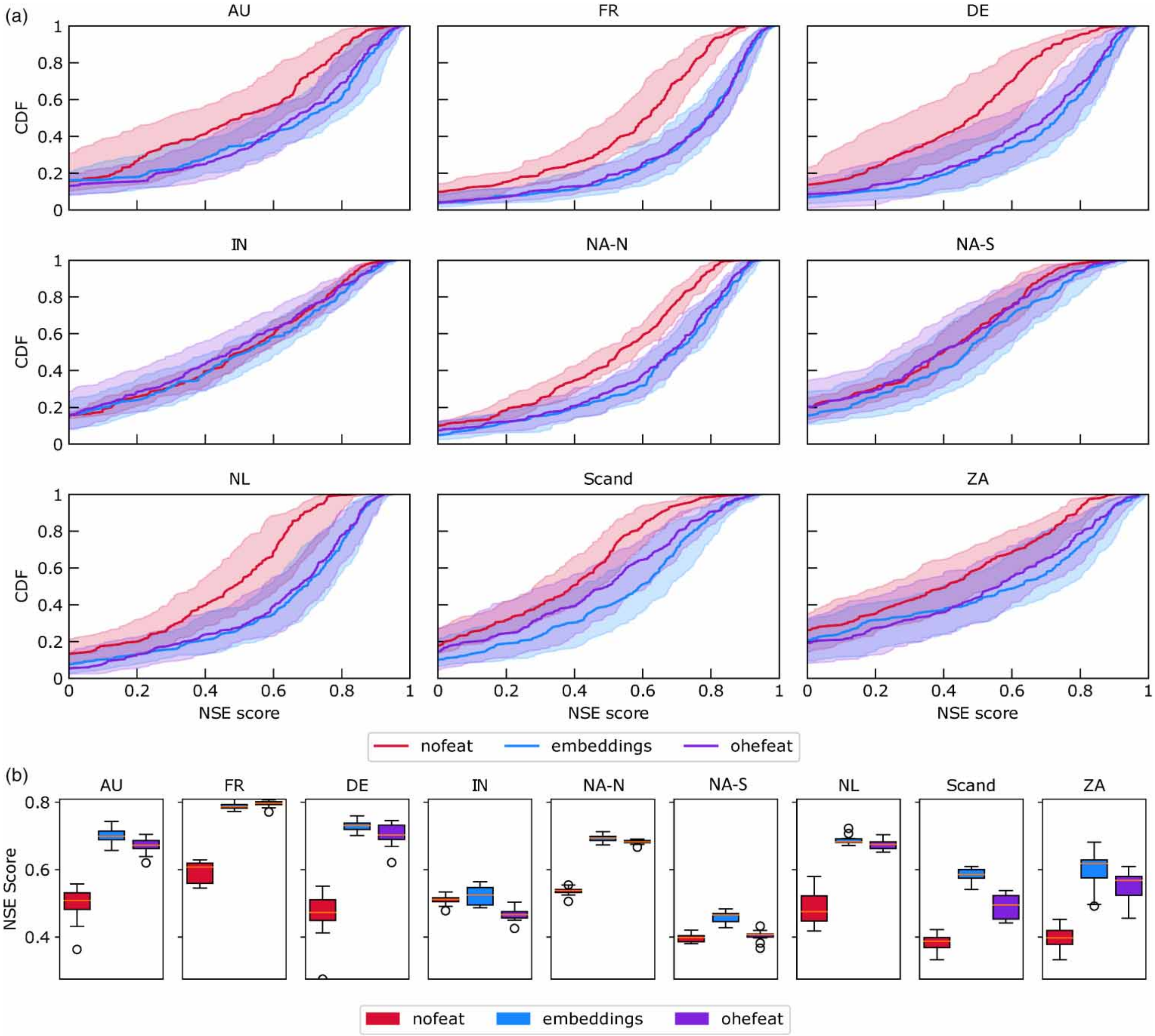


Figure 3 | NSE scores for the three model variants *nofeat*, *embeddings*, and *ohefeat* (Figure 2), across the nine focus regions shown in Figure 1(a). (a) CDFs of individual well NSE scores. Solid lines represent the median across 10 models with different random initializations; shaded envelopes indicate the range of those runs. (b) Boxplots showing the distribution of median NSE scores from the 10 model runs for each region and model variant.

4.3. Global models match regional performance

To evaluate the effects of combining data from regions with richer, higher-quality GWL time series and those with less comprehensive or lower-quality data (compare Section 4.4), we conducted experiments using the entire global GWL dataset, which consists of 1,800 time series. In contrast to the regional model runs discussed in Sections 4.1 and 4.2 (Figure 3 and Table 2), where models were trained separately for each region, the global model is trained on the complete, worldwide dataset, bringing together all available wells into a single training framework. Given the high computational demands of the *ohefeat* model, which become increasingly prohibitive as the number of time series increases, and the superior performance of the *embeddings* model (Section 4.2), we focused our comparison on the *embeddings* and *nofeat* models.

Figure 4 compares the performance of models trained on the full global dataset with those trained separately on each region. In Figure 4(a), the CDFs of NSE scores show that both the global *embeddings* model and the *nofeat* model perform almost identically to their respective ensembles of regional models. The median NSE scores from the global dataset are 0.44 (*nofeat* model) and 0.63 (*embeddings* model), closely aligning with the median across all regional models combined (compare Section 4.1). This consistency also holds at the level of individual wells: as shown in Figure 4(b), most points lie close to the 1:1 line, indicating that NSE values for individual wells remain largely stable when switching from regional to global training. These results have two key implications.

First, increasing the size of the training dataset introduces greater variability, which in theory can help DL models generalize better by exposing them to a broader range of patterns. However, this does not automatically improve model performance for region-specific or local GWL predictions. Previous studies have shown that LSTM models trained in multi-site setups often outperform single-site models (e.g., in Heudorfer *et al.* 2024). Furthermore, training on data-rich regions has been found to improve runoff estimations (Ma *et al.* 2021; Le *et al.* 2024) and GWL simulations (Clark *et al.* 2022) in data-sparse regions. However, unlike previous approaches that randomly combined data or grouped catchments based on hydrological similarity (Fang *et al.* 2022; Kratzert *et al.* 2024) or hydrogeological similarity to reduce data complexity (Chidepudi *et al.* 2025), our regional datasets were defined by geographic proximity. This means that each regional dataset contains a full array of time series from its respective geographic area, capturing a wide range of GWL dynamics that can vary significantly. This variation often occurs even within small spatial extents due to subsurface heterogeneity and scale-dependent processes (Blöschl & Sivapalan 1995; Neuman & Di Federico 2003; Nolte *et al.* 2024). Therefore, the most relevant dynamics for each region (including both dominant and more localized temporal patterns) are likely already well represented within its dataset. We conclude that incorporating additional data from geographically and climatologically dissimilar regions (Supplementary Figure S2) may not provide meaningful context for supporting the modeling task in this study. This is not

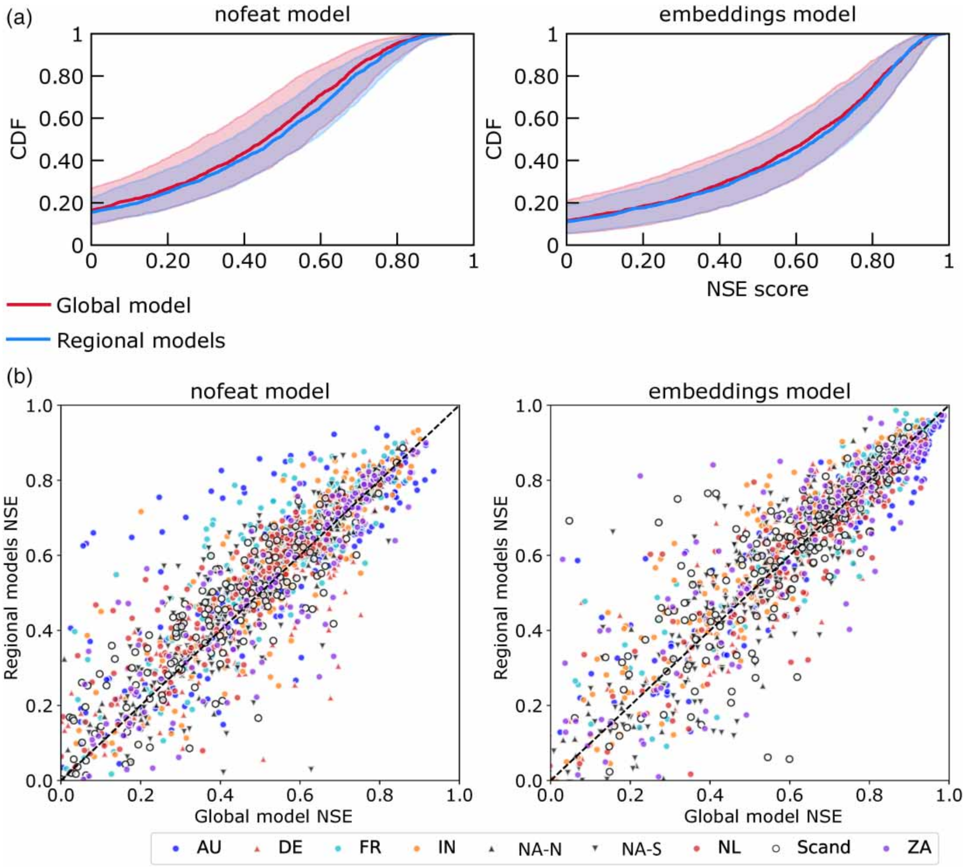


Figure 4 | Comparison of NSE scores for the *nofeat* and *embeddings* models trained on the global dataset (global model) versus models trained separately on regional datasets (regional models). (a) CDFs of individual well NSE scores. Solid lines represent the median across 10 model initializations; shaded envelopes indicate the range across runs. (b) Scatter plots comparing NSE scores from global models (*x*-axis) against regional models (*y*-axis) for individual wells, colored by focus region as defined in Figure 1(a).

because the model cannot integrate heterogeneous inputs, but because the added data do not contribute temporal or hydrogeological patterns beyond those already represented in the target region.

Second, while no additional benefits were found with the global model for the data-sparse regions in this study, the comparable performance of the global model to regional models highlights the potential for multi-site architectures to integrate diverse datasets from around the world without compromising prediction accuracy for individual regions or time series. The ability to effectively incorporate data from regions with varying hydrogeological conditions and data quality reduces the need for time-consuming time series selection. Furthermore, it underscores the potential scalability of these approaches with DL.

While the global model did not yield performance gains for data-sparse regions in this study, previous research has shown that comprehensive datasets can offer substantial benefits under different conditions. For example, transfer learning – where models are applied across regions (Wagenaar *et al.* 2018; Ma *et al.* 2021), typically supported by static input features that help guide cross-regional generalization – has demonstrated such benefits. In contrast, our setup omits such features, relying solely on dynamic inputs and learned site representations. Future experiments could build on this work by leveraging high-resolution datasets (e.g., daily) to improve model performance or simulate finer-scale dynamics in regions where only lower-resolution data (e.g., monthly) are available. Similar advantages have also been shown for addressing unforeseen climatic conditions not represented in the training data (Fang *et al.* 2022; Wi & Steinschneider 2024). Developing such capabilities is critical for advancing DL models toward practical deployment in operational GWL forecasting, though such experiments lie beyond the scope of this study. Overall, the multi-site framework presented here offers a promising foundation for these future directions by enabling modeling across diverse hydrogeological and climatic conditions – even at the global scale and without relying on predefined static features.

4.4. Data quantity and quality limitations

The observed differences in performance across regional models are closely linked to data quantity (the availability of GWL measurements within each time series; Supplementary Figure S1). Figure 5 illustrates the differences by showcasing selected time series whose NSE scores are near the median for their respective region. In general, the model performance is much better in regions with greater data availability and fewer gaps (DE, FR, NL) compared with those with more limited or irregular data (IN, ZA). These findings further support that sufficient and relevant training data improve model performance – not only in single-well data-driven modeling (Wunsch *et al.* 2021) but also in multi-site configurations. In LSTM models, low data availability or quality may hinder stabilization during training, reducing predictive accuracy and increasing variability across runs. As discussed in Section 4.2 for the IN dataset, it appears that there are thresholds in data quantity below which the model struggles to learn and represent GWL dynamics effectively. These thresholds remain understudied in DL groundwater modeling, as most prior studies focused on high-quality, complete time series.

Our results suggest that, beyond the general principle that more information creates better results, a minimum aggregation frequency of monthly observations appears necessary not only for reliable model performance but also for enabling the model to learn site-specific groundwater dynamics. This is evident from the IN models, which were the only ones among all regions to show minimal benefit from including site identifiers, likely due to the very low measurement frequency, with only a few observations per year. Site-specific characteristics, such as how aquifers respond to climatic inputs, are difficult to capture with such sparse data. In addition, models in FR performed well despite being trained on fewer than seven years of data, suggesting that measurement frequency may be more critical for model performance than the total length of the time series. High-frequency inputs allow models to better capture temporal variability, especially in systems influenced by strong seasonal or short-term fluctuations (Taylor & Alley 2001) and to avoid aliasing effects (Bender 2007). Similarly, Wunsch *et al.* (2023) showed that high-frequency (weekly) inputs are essential for learning seasonal and short-term groundwater processes, which are often lost at coarser temporal resolutions.

By contrast, data with longer but consistent intervals smooth out variability, simplifying the prediction task. This effect is seen in the results for IN, where quarterly sampling intervals produced performance comparable to that of regions with irregular, sparse data (ZA, Scand). The irregular intervals and large gaps in these datasets make it more challenging for models to capture and adapt to inconsistent GWL dynamics – even though those dynamics might be equally or more predictable than in IN if the prediction targets were similarly aggregated (quarterly). In this sense, the seemingly better performance in IN may reflect a less complex prediction target, rather than inherently more predictable groundwater behavior. While time series length was less critical for our short-term, weekly-ahead predictions, it becomes essential for long-term groundwater forecasting. Climate-driven GWL trends often evolve slowly and are masked by low-frequency variability, requiring multi-decadal

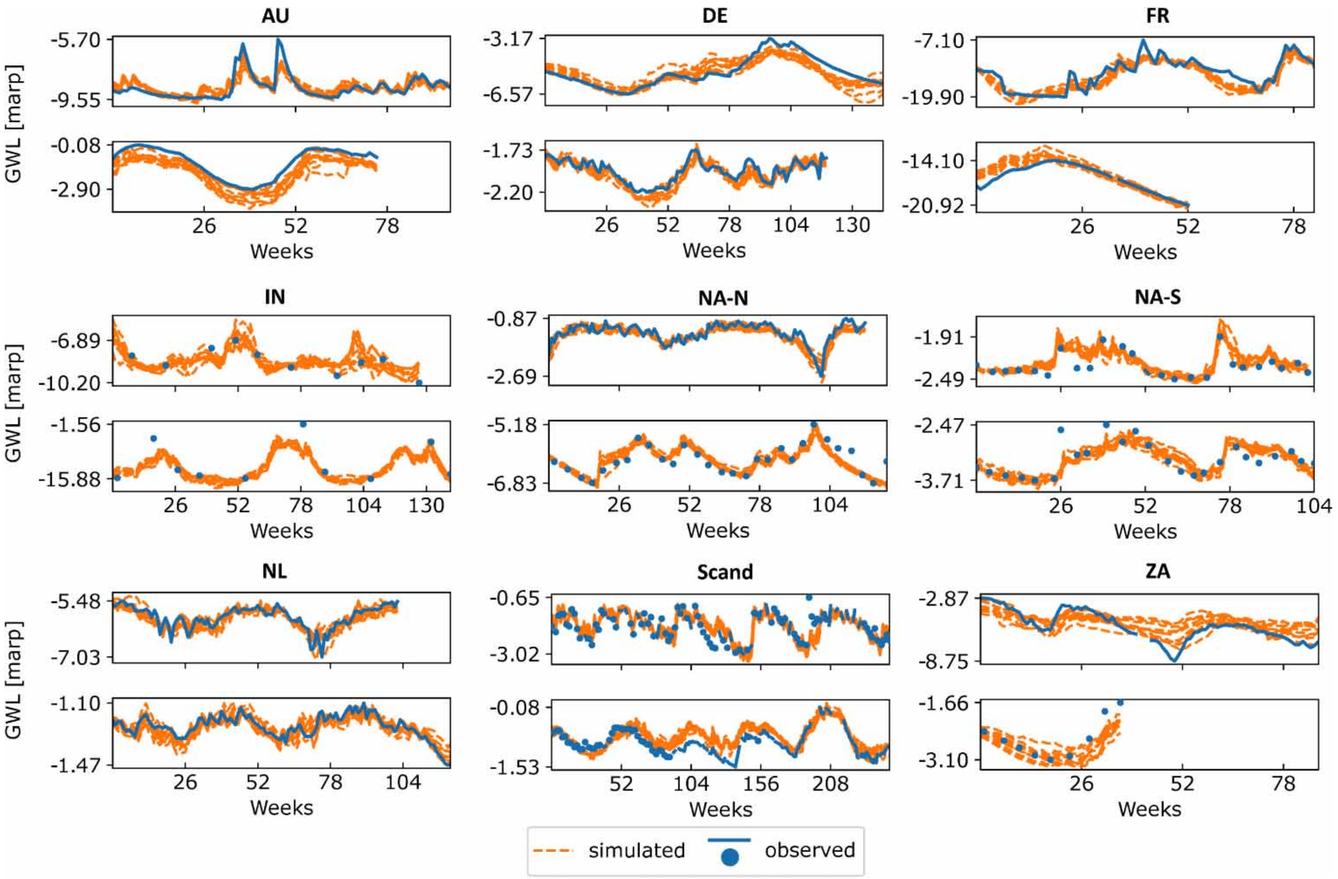


Figure 5 | Example GWL time series in the test period from two wells in each of the nine focus regions (see Figure 1(a)), selected to approximate the median NSE value of their respective regional *embeddings* models. GWL is shown in meters above reference point (marp). Simulated values (orange dashed) are compared against observed measurements (blue). The figure illustrates regional differences in data resolution and prediction accuracy.

records to distinguish persistent changes (Baulon *et al.* 2022), which long-term (DL) forecasts typically rely on (e.g., Wunsch *et al.* 2022). These aspects highlight the importance of aligning the prediction target with data properties when selecting GWL time series and evaluating model performance.

Other factors beyond data quantity that might influence model performance include the complexity of the system being modeled and the quality of the time series data. The characteristics of the modeled aquifer systems, such as groundwater depth, permeability, and aquifer type, can affect how aquifers respond to climatic inputs. While the regional datasets in this study are comparable in terms of groundwater depths (Table 1) and associated climate–groundwater interactions (Cuthbert *et al.* 2019a), detailed information on other aquifer properties is not consistently available across datasets, limiting a more in-depth assessment. Instead, the differing model performances between NA-N and NA-S, despite similar data quantities, may be partially explained by large-scale map products: wells in NA-N are often located in smaller, localized aquifers that may respond rapidly to rainfall, whereas those in NA-S are more often associated with larger regional groundwater systems (Richts *et al.* 2011). Aquifer responsiveness may significantly influence how effectively LSTM models can capture groundwater dynamics in response to changes in precipitation and temperature. Additionally, human activities – such as extensive unsustainable groundwater pumping, as observed in California (de Graaf *et al.* 2019; Liu *et al.* 2022) – can decouple GWLs from climate drivers, thereby reducing model performance. Similar effects have been observed in single-well regional modeling in Northern Germany, where lower performance was attributed to anthropogenic influences like proximity to waterworks (Gomez *et al.* 2024).

Lastly, the accuracy of the dynamic input data must be considered. In this study, inputs were solely derived from ERA5 reanalysis, which offers global consistency but may not match the resolution or precision of regionally available climate data. In some regions or at certain well sites, finer-resolution or more localized inputs might yield locally improved predictions.

4.5. Insights into spatial representation through embeddings

The embedding layer introduced in our LSTM model (see Section 3.3) enables exploration of how GWL dynamics are internally represented. By transforming well identifiers into trainable vectors, the model learns to differentiate between wells in a way that supports GWL prediction from climate inputs. This differentiation is learned directly from the time series data, without relying on externally provided static features. In the following, we examine whether the resulting embedding space may capture meaningful variation across wells that relates to underlying physical conditions.

Despite the inherently abstract nature of embeddings, we found discernible patterns in the reduced embedding space using PCA and t-SNE. Wells grouped by continent (Figure 6(a)) and by country within Europe (Figure 6(b)) show considerable overlap in the embedding space, supporting the findings from Section 4.3, where we hypothesized that the global model does not outperform regional models because a broad variety of GWL dynamics is already represented within regional datasets. Notably, wells from Europe (EU; blue dots), which dominate the dataset, display the widest dispersion in the embedding space, reflecting the high diversity in GWL dynamics across the continent. In contrast, wells from India (IN; black dots) cluster more tightly, suggesting more homogeneous behavior among the 200 Indian wells compared with the 1,000 wells from different parts of Europe. Among the visible groupings in the 2D t-SNE space, distinct subclusters appear, including those in the right corner containing wells from the Netherlands (blue dots in Figure 6(a) and 6(b)) and Australia (turquoise dots in Figure 6(a)). The partial alignment of embeddings with geographic groupings is expected to some extent, given similarities in climate-driven GWL responses, as also suggested by the performance of climate-only models in Section 4.2. At the same time, the observed differences in overlap and separation between and within regions suggest that the embeddings may also capture site-specific functional variation, such as differences in hydrograph behavior.

To test this, we compared embedding-based clustering to hydrograph-based GWL clusters (see Section 3.4), which represent distinct temporal response patterns such as differences in seasonal amplitude, timing, and regularity (Nolte *et al.* 2024). The alignment between both clusterings was moderate, with silhouette scores between 0.17 and 0.21 and NMI scores between 0.24 and 0.31 across runs. While the relatively low silhouette scores indicate weak to moderate separation (Kaufman & Rousseeuw 1990), this is consistent with the original hydrograph-based clusters, which were not sharply delineated. The NMI scores offer further context: they are comparable to values reported in other hydrological studies, such as Gupta & Karthikeyan (2024), where similarly modest NMI scores emerged when comparing clusters based on different drought-related features. In that context, the low agreement reflected both the fuzzy nature of functional types and the fact that different input features emphasize different aspects of system behavior. Likewise, in our case, embedding-based and hydrograph-based clusterings offer distinct but partially overlapping perspectives on groundwater dynamics.

The partial alignment with hydrograph types, together with the visual patterns in Figure 7, suggests that the embedding space captures meaningful variation across wells related to underlying physical conditions such as climate, water table depth, and lithology (Nolte *et al.* 2024). In particular, the hydrograph-based clusters can be visually distinguished in the embedding space, forming partially separated regions that support the idea that embeddings reflect functional similarity across wells. This occurred despite the model not receiving any explicit physical descriptors, highlighting the potential of embeddings to differentiate groundwater behavior beyond simple geographic correlation and to infer hydrogeological similarity directly from time series alone (Barthel *et al.* 2021).

While this suggests encouraging potential, the low silhouette and NMI values also underline a fundamental limitation: statistical clustering may struggle to capture functional (dis)similarity in groundwater systems, where dynamics often form gradual continua rather than discrete classes. Confirming hydrogeological similarity in embedding space, therefore, remains challenging without corresponding high-resolution or even site-specific environmental data. Future work should therefore aim to validate embedding interpretations using independent descriptors where available. Yet, recent studies highlight the limitations of this approach in practice: Chidepudi *et al.* (2025), working with wells across France, found that static environmental features often lack sufficient resolution or reliability to support meaningful interpretation using tools like SHAP, while Kunz *et al.* (2024), in a Germany-based study, observed only marginal performance gains from including static descriptors and noted that many such features fail to capture site-specific hydrogeological conditions, thus constraining both predictive value and interpretability.

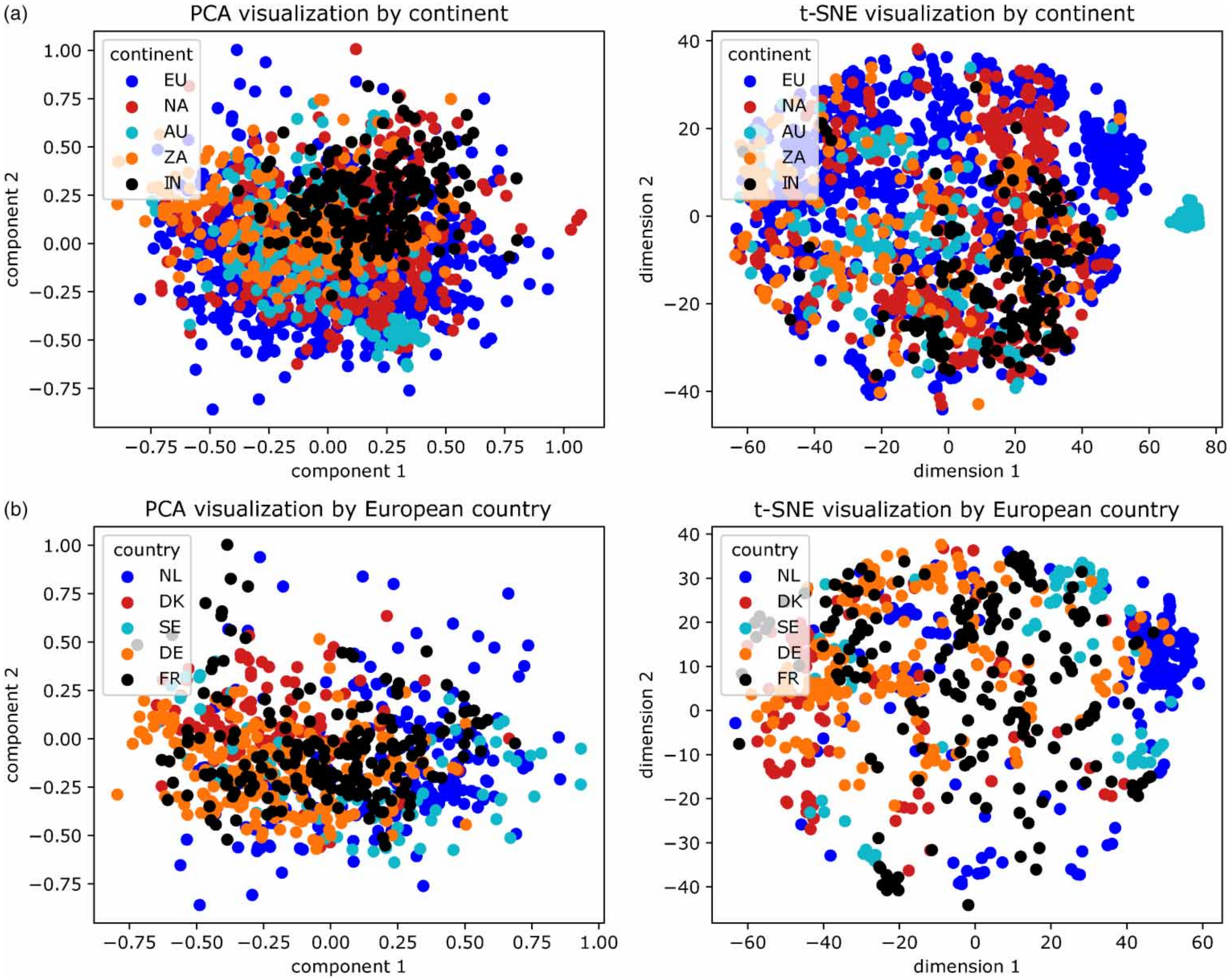


Figure 6 | Visualization of the embedding space from the global model using PCA and t-SNE, colored by (a) continent such as Europe (EU), North America (NA), Australia (AU), South Africa (ZA), and India (IN) and (b) European country such as the Netherlands (NL), Denmark (DK), Sweden (SE), Germany (DE), and France (FR). Each point represents a well, and its position reflects relationships in GWL dynamics as learned by the model through embeddings.

The interpretability of embeddings depends not only on actual groundwater behavior but also on data quality. Sparse, short, irregular, or noisy time series can distort embedding formation, not due to hydrogeological distinctiveness, but because of reduced learnability. This introduces a circular dependency: poor data quality can degrade embeddings, which in turn limits interpretability and obscures the underlying cause. To investigate this, we applied DBSCAN (see Section 3.4) to flag outliers in the embedding space and compared their predictive performance. Across regional models, these outlier wells consistently showed lower NSE scores than inliers (Figure 8), suggesting that they are either underrepresented in the training data or exhibit atypical behavior due to natural variability, anthropogenic influence, or poor data quality.

Looking ahead in terms of embedding implementation, evaluating embedding stability across model runs could provide deeper insight into the robustness of learned representations. Similarly, testing lower-dimensional embeddings, such as two or three dimensions, may offer more interpretable representations without loss of predictive performance. Preliminary tests (not shown) indicated that reducing from seven to three dimensions preserved nearly identical predictive skill, suggesting that compact representations (Bassi *et al.* 2024) may be sufficient for both modeling and interpretation.

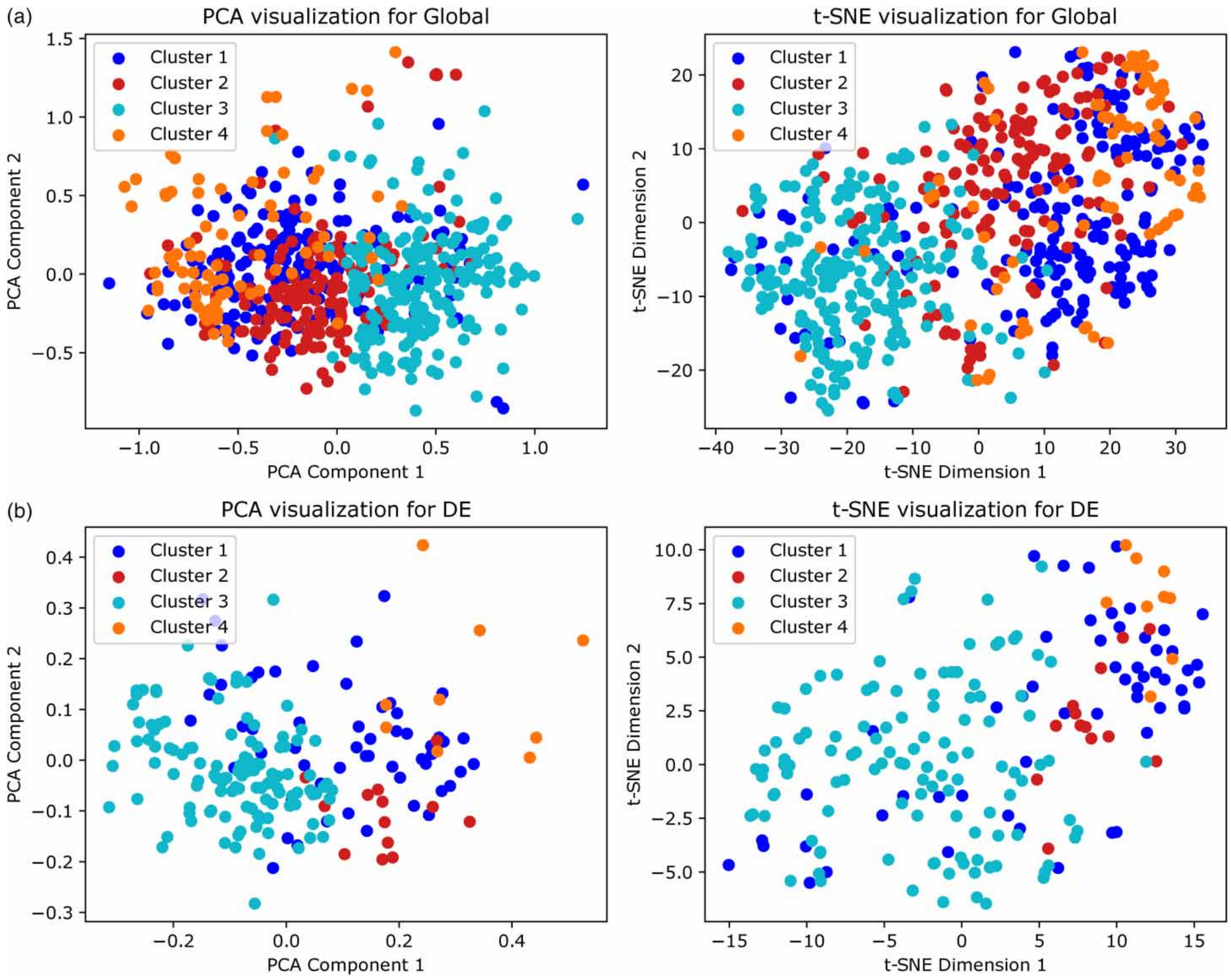


Figure 7 | Visualization of the embedding space from (a) the global model and (b) the DE model using PCA and t-SNE. Wells that are shared between this study and Nolte *et al.* (2024) are plotted and color-coded according to four clusters representing distinct GWL dynamics, as previously defined using hydrograph analysis (Nolte *et al.* 2024). Each point represents a well, and its position reflects relationships in GWL dynamics as learned by the model through embeddings.

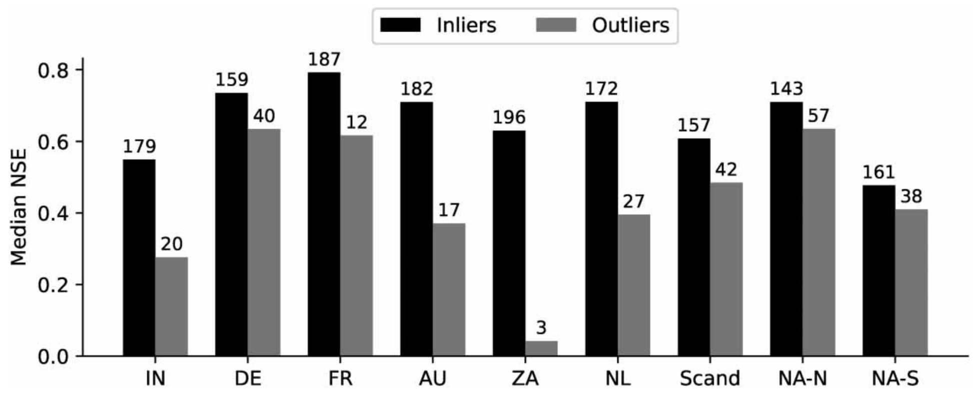


Figure 8 | Median NSE scores for inlier and outlier wells across nine focus regions, based on DBSCAN clustering (with eps = 0.25, min_-samples = 2) in the embedding space. Numbers above bars indicate the number of wells per group (inliers in black, outliers in gray). The comparison reveals a systematically lower model performance for wells identified as outliers in the embedding space.

5. CONCLUSIONS

This study advances large-scale GWL modeling in three key ways. First, it presents the first global, systematic evaluation of multi-site LSTM models across more than 1,800 globally distributed wells, using GWL time series with realistic variability in completeness, sampling frequency, and potential noise. Second, it demonstrates that trainable site identification (also called embeddings) can partially and effectively substitute for static environmental descriptors, which are often unavailable or uncertain, and which recent studies have shown to function more as unique site identifiers than physically meaningful predictors. Third, it provides the first analysis of the learned embedding space in a groundwater context, revealing emergent spatial and functional patterns that suggest the model captures some relevant hydrogeological structure even without explicit physical input. More broadly, the methods explored here may also support applications across the wider field of hydrology.

Embeddings provide a scalable and efficient alternative to one-hot encodings. Embedding-based models achieved high performance in data-rich regions, with median NSE scores exceeding 0.7 in countries like France and Germany. This indicates that the models not only work technically but also achieve a level of predictive accuracy that allows for practical application in regional analysis where data availability is sufficient. The analysis of the embedding space suggests that DL models can internally capture site-specific behavior directly from time series data, which, in the case of GWL time series, can be related to differences in seasonal recharge dynamics, response timing, or longer-term variability – all of which may reflect underlying factors such as lithology or water table depth. The embeddings approach, therefore, holds promise not only for improving predictive performance but also for uncovering previously unrecognized patterns in groundwater systems, opening new ways for data-driven discovery. However, confidence in explaining these representations remains limited because this ultimately depends on the availability of independent physical descriptors, which may be explored through targeted case studies where such data exist.

The study also shows that global models perform comparably to regional models, even when integrating lower-quality data. This eliminates the need for separate region-specific or single-site models and highlights the potential for truly geographically global DL applications. However, a limitation is that adding high-quality GWL time series from other regions did not improve predictions in data-sparse regions. Achieving reliable performance in such settings remains a key challenge for multi-site DL models and warrants further investigation. These findings underscore the importance of acquiring sufficient data coverage, particularly at high temporal resolution, within the same regions for which predictive accuracy is currently limited.

Looking ahead, entity-aware multi-site models offer a promising direction for scalable time series prediction at continental to global scales. To support operational use cases such as real-time monitoring and climate impact assessments, future work should evaluate the model performance in true forecasting settings and further assess the stability and interpretability of site embeddings. Particular attention is needed for data-scarce or irregular time series, where strategies like transfer learning may help overcome current limitations in model accuracy. While embeddings cannot generalize to entirely unseen wells, they provide a powerful solution when static input features are missing or unreliable, and may be extended to new wells through targeted fine-tuning. Future work may also explore more advanced data fusion strategies beyond simple concatenation to better integrate static and dynamic information. Together, these directions will help advance DL toward becoming a practical and reliable tool for large-scale groundwater prediction under real-world data constraints.

FUNDING

This work was financed within the framework of the Helmholtz Institute for Climate Service Science (HICSS), cooperation between the Climate Service Center Germany (GERICS) and Universität Hamburg, Germany, and conducted as part of the Future-H_2O (Future climate and land-use change impacts on groundwater recharge rates and water quality for water resources) project.

AUTHOR CONTRIBUTIONS

A.N.: Conceptualization, Data curation, Methodology, Software, Formal analysis, Writing – original draft, Writing – review and editing. B.H.: Conceptualization, Methodology, Software, Writing – review and editing. S.B.: Funding acquisition, Supervision, Writing – review and editing. J.H.: Funding acquisition, Supervision, Writing – review and editing.

DATA AVAILABILITY STATEMENT

All relevant data are available from an online repository or repositories: https://doi.org/10.5281/zenodo.14834540.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Addor, N., Newman, A. J., Mizukami, N. & Clark, M. P. (2017) The CAMELS data set: catchment attributes and meteorology for large-sample studies, *Hydrology and Earth System Sciences*, **21** (10), 5293–5313. doi: 10.5194/hess-21-5293-2017.

Akbik, A., Blythe, D. & Vollgraf, R. (2018) Contextual string embeddings for sequence labeling, *Proceedings of the 27th International Conference on Computational Linguistics*. Stroudsburg, PA: Association for Computational Linguistics, pp. 1638–1649.

Balacumaresan, H., Imteaz, M. A., Aziz, M. A. & Choudhury, T. (2024) Use of artificial intelligence modelling for the dynamic simulation of urban catchment runoff, *Water Resources Management*, **38** (10), 3657–3683. doi: 10.1007/s11269-024-03833-9.

Barthel, R., Haaf, E., Giese, M., Nygren, M., Heudorfer, B. & Stahl, K. (2021) Similarity-based approaches in hydrogeology: proposal of a new concept for data-scarce groundwater resource characterization and prediction, *Hydrogeology Journal*, **24** (4), 3392. doi: 10.1007/s10040-021-02358-4.

Barthel, R., Haaf, E., Nygren, M. & Giese, M. (2022) Systematic visual analysis of groundwater hydrographs: potential benefits and challenges, *Hydrogeology Journal*, **30** (2), 359–378. doi: 10.1007/s10040-021-02433-w.

Bassi, A., Höge, M., Mira, A., Fenicia, F. & Albert, C. (2024) Learning landscape features from streamflow with autoencoders, *Hydrology and Earth System Sciences*, **2024**, 1–30. doi: 10.5194/hess-28-4971-2024.

Baulon, L., Massei, N., Allier, D., Fournier, M. & Bessiere, H. (2022) Influence of low-frequency variability on high and low groundwater levels: example of aquifers in the Paris Basin, *Hydrology and Earth System Sciences*, **26** (11), 2829–2854. doi: 10.5194/hess-26-2829-2022.

Bender, S. (2007) Die Aussageunschärfe bei der Verwendung heterogener Datensätze im Rahmen wasserwirtschaftlicher Fragestellungen [The ambiguity of conclusions when using heterogeneous data sets in the context of water management issues], *Bochumer Geowissenschaftliche Arbeiten*, **8**, 99.

Bloomfield, J. P., Marchant, B. P. & McKenzie, A. A. (2019) Changes in groundwater drought associated with anthropogenic warming, *Hydrology and Earth System Sciences*, **23** (3), 1393–1408. doi: 10.5194/hess-23-1393-2019.

Blöschl, G. & Sivapalan, M. (1995) Scale issues in hydrological modelling: a review, *Hydrological Processes*, **9** (3-4), 251–290. doi: 10.1002/hyp.3360090305.

Botterill, T. E. & McMillan, H. K. (2023) Using machine learning to identify hydrologic signatures with an encoder–decoder framework, *Water Resources Research*, **59** (3), 19. doi: 10.1029/2022WR033091.

Chidepudi, S. K. R., Massei, N., Jardani, A., Henriot, A., Allier, D. & Baulon, L. (2023) A wavelet-assisted deep learning approach for simulating groundwater levels affected by low-frequency variability, *Science of the Total Environment*, **865**, 161035. doi: 10.1016/j.scitotenv.2022.161035.

Chidepudi, S. K. R., Massei, N., Jardani, A., Dieppois, B., Henriot, A. & Fournier, M. (2025) 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? *Hydrology and Earth System Sciences*, **2025**, 1–28. doi: 10.5194/hess-29-841-2025.

Clark, S. R., Pagendam, D. & Ryan, L. (2022) Forecasting multiple groundwater time series with local and global deep learning networks, *International Journal of Environmental Research and Public Health*, **19** (9), 31. doi: 10.3390/ijerph19095091.

Cuthbert, M. O., Gleeson, T., Moosdorf, N., Befus, K. M., Schneider, A., Hartmann, J. & Lehner, B. (2019a) Global patterns and dynamics of climate–groundwater interactions, *Nature Climate Change*, **9** (2), 137–141. doi: 10.1038/s41558-018-0386-4.

Cuthbert, M. O., Taylor, R. G., Favreau, G., Todd, M. C., Shamsudduha, M., Villholth, K. G., MacDonald, A. M., Scanlon, B. R., Kotchoni, D. V. & Vouillamoz, J.-M. (2019b) Observed controls on resilience of groundwater to climate variability in sub-Saharan Africa, *Nature*, **572** (7768), 230–234. doi: 10.1038/s41586-019-1441-7.

de Graaf, I., Gleeson, T., van Rens Beek, L. P. H., Sutanudjaja, E. H. & Bierkens, M. F. P. (2019) Environmental flow limits to global groundwater pumping, *Nature*, **574** (7776), 90–94. doi: 10.1038/s41586-019-1594-4.

Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. (1996) A density-based algorithm for discovering clusters in large spatial databases with noise, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, Vol. 34. Washington, DC: AAAI Press, pp. 226–231.

Fang, K., Kifer, D., Lawson, K., Feng, D. & Shen, C. (2022) The data synergy effects of time-series deep learning models in hydrology, *Water Resources Research*, **58** (4), e2021WR029583. doi: 10.1029/2021WR029583.

Frame, J. M., Kratzert, F., Klotz, D., Gauch, M., Shalev, G., Gilon, O., Qualls, L. M., Gupta, H. V. & Nearing, G. S. (2022) Deep learning rainfall–runoff predictions of extreme events, *Hydrology and Earth System Sciences*, **26** (13), 3377–3392. doi: 10.5194/hess-26-3377-2022.

Gleeson, T., Cuthbert, M., Ferguson, G. & Perrone, D. (2020a) Global groundwater sustainability, resources, and systems in the Anthropocene, *Annual Review of Earth and Planetary Sciences*, **48** (1), 431–463. doi: 10.1146/annurev-earth-071719-055251.

Gleeson, T., Wang-Erlandsson, L., Porkka, M., Zipper, S. C., Jaramillo, F., Gerten, D., Fetzer, I., Cornell, S. E., Piemontese, L. & Gordon, L. J. (2020b) Illuminating water cycle modifications and Earth system resilience in the Anthropocene, *Water Resources Research*, **56** (4), e2019WR024957. doi: 10.1029/2019WR024957.

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., Graaf, I. d., Cuthbert, M., Condon, L., Bresciani, E. & Bierkens, M. F. P. (2021) GMD perspective: the quest to improve the evaluation of groundwater representation in continental- to global-scale models, *Geoscientific Model Development*, **14** (12), 7545–7571. doi: 10.5194/gmd-14-7545-2021.

Gomez, M., Nölscher, M., Hartmann, A. & Broda, S. (2024) Assessing groundwater level modelling using a 1-D convolutional neural network (CNN): linking model performances to geospatial and time series features, *Hydrology and Earth System Sciences*, **28** (19), 4407–4425. doi: 10.5194/hess-28-4407-2024.

Gupta, A. & Karthikeyan, L. (2024) Role of initial conditions and meteorological drought in soil moisture drought propagation: an event-based causal analysis over South Asia, *Earth's Future*, **12** (10), e2024EF004674. doi: 10.1029/2024EF004674.

Haaf, E., Giese, M., Reimann, T. & Barthel, R. (2023) Data-driven estimation of groundwater level time-series at unmonitored sites using comparative regional analysis, *Water Resources Research*, **59**, e2022WR033470. doi: 10.1029/2022WR033470.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R. & Schepers, D. (2020) The ERA5 global reanalysis, *Quarterly Journal of the Royal Meteorological Society*, **146** (730), 1999–2049. doi: 10.1002/qj.3803.

Heudorfer, B., Haaf, E., Stahl, K. & Barthel, R. (2019) Index-based characterization and quantification of groundwater dynamics, *Water Resources Research*, **55** (7), 5575–5592. doi: 10.1029/2018WR024418.

Heudorfer, B., Liesch, T. & Broda, S. (2024) On the challenges of global entity-aware deep learning models for groundwater level prediction, *Hydrology and Earth System Sciences*, **28** (3), 525–543. doi: 10.5194/hess-28-525-2024.

Heudorfer, B., Gupta, H. V. & Loritz, R. (2025) Are deep learning models in hydrology entity aware? *Geophysical Research Letters*, **52** (6), e2024GL113036. doi: 10.1029/2024GL113036.

Hochreiter, S. & Schmidhuber, J. (1997) Long short-term memory, *Neural Computation*, **9** (8), 1735–1780. doi: 10.1162/neco.1997.9.8.1735.

Kaufman, L. & Rousseeuw, P. J. (2009) *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.

Kingma, D. P. & Ba, J. L. (2014) Adam: a method for stochastic optimization, *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. San Diego, CA: ICLR.

Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S. & Nearing, G. (2022) Uncertainty estimation with deep learning for rainfall–runoff modeling, *Hydrology and Earth System Sciences*, **26** (6), 1673–1693. doi: 10.5194/hess-26-1673-2022.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K. & Herrnegger, M. (2018) Rainfall–runoff modelling using long short-term memory (LSTM) networks, *Hydrology and Earth System Sciences*, **22** (11), 6005–6022. doi: 10.5194/hess-22-6005-2018.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S. & Nearing, G. (2019) Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, *Hydrology and Earth System Sciences*, **23** (12), 5089–5110. doi: 10.5194/hess-23-5089-2019.

Kratzert, F., Gauch, M., Klotz, D. & Nearing, G. (2024) HESS opinions: never train a long short-term memory (LSTM) network on a single basin, *Hydrology and Earth System Sciences*, **28** (17), 4187–4201. doi: 10.5194/hess-28-4187-2024.

Kulp, S. A. & Strauss, B. H. (2019) New elevation data triple estimates of global vulnerability to sea-level rise and coastal flooding, *Nature Communications*, **10** (1), 1–12. doi: 10.1038/s41467-019-12808-z.

Kunz, S., Schulz, A., Wetzel, M., Nölscher, M., Chiaburu, T., Biessmann, F. & Broda, S. (2024) Towards a global spatial machine learning model for seasonal groundwater level predictions in Germany, *EGUsphere* (Submitted), **2024**, 1–44. doi: 10.5194/egusphere-2024-3484.

Lall, U., Josset, L. & Russo, T. (2020) A snapshot of the world's groundwater challenges, *Annual Review of Environment and Resources*, **45** (1), 171–194. doi: 10.1146/annurev-environ-102017-025800.

Le, M.-H., Kim, H., Do, H. X., Beling, P. A. & Lakshmi, V. (2024) A framework on utilizing of publicly availability stream gauges datasets and deep learning in estimating monthly basin-scale runoff in ungauged regions, *Advances in Water Resources*, **188**, 104694. doi: 10.1016/j.advwatres.2024.104694.

Lee, S., Lee, K.-K. & Yoon, H. (2019) Using artificial neural network models for groundwater level forecasting and assessment of the relative impacts of influencing factors, *Hydrogeology Journal*, **27** (2), 567–579. doi: 10.1007/s10040-018-1866-3.

Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G. & Dadson, S. J. (2021) Benchmarking data-driven rainfall–runoff models in Great Britain: a comparison of long short-term memory (LSTM)-based models with four lumped conceptual models, *Hydrology and Earth System Sciences*, **25** (10), 5517–5534. doi: 10.5194/hess-25-5517-2021.

Li, X., Khandelwal, A., Jia, X., Cutler, K., Ghosh, R., Renganathan, A., Xu, S., Tayal, K., Nieber, J. & Duffy, C. (2022) Regionalization in a global hydrologic deep learning model: from physical descriptors to random vectors, *Water Resources Research*, **58** (8), e2021WR031794. doi: 10.1029/2021WR031794.

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. & Rodell, M. (2022) Groundwater depletion in California's Central Valley accelerates during megadrought, *Nature Communications*, **13** (1), 7825. doi: 10.1038/s41467-022-35582-x.

Loritz, R., Wu, C. H., Klotz, D., Gauch, M., Kratzert, F. & Bassiouni, M. (2024) Generalizing tree-level sap flow across the European continent, *Geophysical Research Letters*, **51** (6), e2023GL107350. doi: 10.1029/2023GL107350.

Ma, K., Feng, D., Lawson, K., Tsai, W.-P., Liang, C., Huang, X., Sharma, A. & Shen, C. (2021) Transferring hydrologic data across continents–leveraging data-rich regions to improve hydrologic prediction in data-sparse regions, *Water Resources Research*, **57** (5), e2020WR028600. doi: 10.1029/2020WR028600.

Mangukiya, N. K. & Sharma, A. (2025) Deep learning-based approach for enhancing streamflow prediction in watersheds with aggregated and intermittent observations, *Water Resources Research*, **61** (1), e2024WR037331. doi: 10.1029/2024WR037331.

Massei, N., Kingston, D. G., Hannah, D. M., Vidal, J.-P., Dieppois, B., Fossa, M., Hartmann, A., Lavers, D. A. & Laignel, B. (2020) Understanding and predicting large-scale hydrological variability in a changing environment, *Proceedings of the International Association of Hydrological Sciences*, **383**, 141–149. doi: 10.5194/piahs-383-141-2020.

Mekanik, F., Imteaz, M. A., Gato-Trinidad, S. & Elmahdi, A. (2013) Multiple regression and artificial neural network for long-term rainfall forecasting using large scale climate modes, *Journal of Hydrology*, **503**, 11–21. doi: 10.1016/j.jhydrol.2013.08.035.

Mohammadi, M., Gato-Trinidad, S. & King Kuok, K. (2024) The limitation of machine learning methods for water supply and demand forecasting: a case study for Greater Melbourne, Australia, *Water Supply*, **24** (11), 3848–3861. doi: 10.2166/ws.2024.225.

Najah, A., El-Shafie, A., Karim, O. A. & El-Shafie, A. H. (2013) Application of artificial neural networks for water quality prediction, *Neural Computing and Applications*, **22** (Suppl. 1), 187–201.

Neuman, S. P. & Di Federico, V. (2003) Multifaceted nature of hydrogeologic scaling and its interpretation, *Reviews of Geophysics*, **41** (3), 1014. doi: 10.1029/2003RG000130.

Nolte, A., Haaf, E., Heudorfer, B., Bender, S. & Hartmann, J. (2024) Disentangling coastal groundwater level dynamics in a global dataset, *Hydrology and Earth System Sciences*, **28** (5), 1215–1249. doi: 10.5194/hess-28-1215-2024.

Rajaee, T., Ebrahimi, H. & Nourani, V. (2019) A review of the artificial intelligence methods in groundwater level modeling, *Journal of Hydrology*, **572**, 336–351. doi: 10.1016/j.jhydrol.2018.12.037.

Retike, I., Bikše, J., Kalvāns, A., Dēliņa, A., Avotniece, Z., Zaadnoordijk, W. J., Jemeljanova, M., Popovs, K., Babre, A. & Zelenkevičs, A. (2022) Rescue of groundwater level time series: how to visually identify and treat errors, *Journal of Hydrology*, **605**, 127294. doi: 10.1016/j.jhydrol.2021.127294.

Richts, A., Struckmeier, W. F. & Zaepke, M., (2011) WHYMAP and the groundwater resources map of the world 1:25,000,000. In: Jones, J. A. A. (ed.) *Sustaining Groundwater Resources*. Dordrecht: Springer Netherlands, pp. 159–173.

Rousseeuw, P. J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, *Journal of Computational and Applied Mathematics*, **20**, 53–65. doi: 10.1016/0377-0427(87)90125-7.

Sanford, W. (2002) Recharge and groundwater models: an overview, *Hydrogeology Journal*, **10** (1), 110–120. doi: 10.1007/s10040-001-0173-5.

Scanlon, B. R., Healy, R. W. & Cook, P. G. (2002) Choosing appropriate techniques for quantifying groundwater recharge, *Hydrogeology Journal*, **10** (1), 18–39. doi: 10.1007/s10040-001-0176-2.

Shen, C. & Lawson, K. (2021) Applications of deep learning in hydrology. In: Camps-Valls, G., Tuia, D., Zhu, X. X. & Reichstein, M. (eds.). *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences*. Wiley Online Library, pp. 283–297. doi: 10.1002/9781119646181.ch19.pp.

Solgi, R., Loáiciga, H. A. & Kram, M. (2021) Long short-term memory neural network (LSTM-NN) for aquifer level time series forecasting using in-situ piezometric observations, *Journal of Hydrology*, **601**, 126800. doi: 10.1016/j.jhydrol.2021.126800.

Strehl, A. & Ghosh, J. (2002) Cluster ensembles – a knowledge reuse framework for combining multiple partitions, *Journal of Machine Learning Research*, **3** (Dec), 583–617.

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. (2022) Groundwater level prediction using machine learning models: a comprehensive review, *Neurocomputing*, **489**, 271–308. doi: 10.1016/j.neucom.2022.03.014.

Taylor, C. J. & Alley, W. M. (2001) *Groundwater-Level Monitoring and the Importance of Long-Term Water-Level Data*. Denver, CO, USA: US Geological Survey.

United Nations (2022) *The United Nations World Water Development Report 2022: Groundwater – Making the Invisible Visible*. Paris: UNESCO (on behalf of UN-Water), p. 246.

van der Maaten, L. & Hinton, G. (2008) Visualizing data using t-SNE, *Journal of Machine Learning Research*, **9** (11), 2579–2605.

Wagenaar, D., Lüdtke, S., Schröter, K., Bouwer, L. M. & Kreibich, H. (2018) Regional and temporal transferability of multivariable flood damage models, *Water Resources Research*, **54** (5), 3688–3703. doi: 10.1029/2017WR022233.

Wang, S., Zhou, W. & Jiang, C. (2020) A survey of word embeddings based on deep learning, *Computing*, **102** (3), 717–740. doi: 10.1007/s00607-019-00768-7.

Wi, S. & Steinschneider, S. (2024) On the need for physical constraints in deep learning rainfall–runoff projections under climate change: a sensitivity analysis to warming and shifts in potential evapotranspiration, *Hydrology and Earth System Sciences*, **28** (3), 479–503. doi: 10.5194/hess-28-479-2024.

Wunsch, A., Liesch, T. & Broda, S. (2021) Groundwater level forecasting with artificial neural networks: a comparison of long short-term memory (LSTM), convolutional neural networks (CNNs), and non-linear autoregressive networks with exogenous input (NARX), *Hydrology and Earth System Sciences*, **25** (3), 1671–1687. doi: 10.5194/hess-25-1671-2021.

Wunsch, A., Liesch, T. & Broda, S. (2022) Deep learning shows declining groundwater levels in Germany until 2100 due to climate change, *Nature Communications*, **13**, 1–13. doi: 10.1038/s41467-022-28770-2.

Wunsch, A., Liesch, T. & Goldscheider, N. (2023) Towards understanding the influence of seasons on low groundwater periods based on explainable machine learning, *Hydrology and Earth System Sciences*, **2023**, 1–19. doi: 10.5194/hess-28-2167-2024.

First received 18 July 2025; accepted in revised form 2 September 2025. Available online 24 September 2025