https://doi.org/10.5194/egusphere-2025-4055
Preprint. Discussion started: 6 October 2025

Never Train a Deep Learning Model on a Single Well? Revisiting Training Strategies for Groundwater Level Prediction

Marc Ohmer[1] and Tanja Liesch[1]

[1]Institute for Applied Geosciences (AGW), Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

Correspondence: Marc Ohmer (marc.ohmer@kit.edu]

Abstract. Deep learning (DL) models are increasingly used for hydrological forecasting, with a growing shift from site-specific to globally trained architectures. This study tests whether the widely held assumption that global models consistently outperform local ones also applies to groundwater systems, which differ substantially from surface water due to slow response dynamics, data scarcity, and strong site heterogeneity. Using a benchmark dataset of nearly 3,000 monitoring wells across Germany, we systematically compare global Long Short-Term Memory (LSTM) models with locally trained single-well models in terms of overall performance, training data characteristics, prediction of extremes, and spatial generalization.

For groundwater level prediction, we find that global models provide no systematic accuracy advantage over local models. Local models more often capture site-specific behavior, while global models yield more robust but less specialized predictions across diverse wells. Performance gains arise primarily from dynamically coherent training data, whereas random data reduction has little effect, indicating that similarity matters more than quantity in this setting. Both model types struggle with extreme groundwater conditions, and global models generalize reliably only to wells with comparable dynamics.

These findings qualify the assumption of global model superiority and highlight the need to align modeling strategies with groundwater-specific constraints and application goals.

1 Introduction

In recent years, deep learning (DL) has transformed hydrological forecasting, with global models often outperforming site-specific approaches for streamflow prediction. However, whether these advances extend to groundwater remains an open question. Groundwater systems differ fundamentally from surface waters: their responses are slower, more heterogeneous, and supported by much sparser data. This raises a critical question for groundwater research: can globally trained models truly outperform locally trained ones, or do groundwater-specific dynamics favor single-well approaches?

Traditionally, hydrological predictions have relied on physically based, process-oriented models. While powerful, these models demand extensive domain expertise, high-quality input data, and often face considerable implementation challenges, inherent uncertainties, and limited transferability across regions (Nayak et al., 2006). For groundwater, additional hurdles arise from geological complexity and the need for long observation periods supported by costly monitoring infrastructures (Chidepudi et al., 2025).

Against this backdrop, data-driven methods, particularly DL, offer a compelling alternative. These models can learn hydrological relationships directly from data, reducing the need for detailed local information (Hauswirth et al., 2021; Gomez et al., 2024), and efficiently capture nonlinear, time-lagged dependencies that characterize systems with strong storage effects such as groundwater, soil moisture, or snowmelt processes (Kratzert et al., 2019; Clark et al., 2022). Common DL architectures include recurrent neural networks (RNNs) such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), as well as convolutional neural networks (CNNs).

Driven by the increasing availability of hydrological data and advances in machine learning, modeling strategies have shifted from locally calibrated, site-specific approaches toward regional and global architectures trained on data from many wells simultaneously, with the aim of extracting generalizable patterns from distributed time series (Nearing et al., 2021; Kratzert et al., 2024). Current DL strategies in hydro(geo)logy can thus be broadly categorized into two types: local (single-well) models and global models, the latter sometimes further refined into partitioned subsets motivated by spatial or dynamic similarity.

Local Models (Single-Well/-Basin Models)

Single-well models are trained individually for each monitoring site, based on the assumption that each time series originates from its own data-generating process (Clark et al., 2022). These models enable a detailed representation of local hydrogeological characteristics and dynamic changes (Wunsch et al., 2022a; Zhang et al., 2025), offering high interpretability due to their sensitivity to site-specific input features and time windows. However, their applicability is primarily limited by a tendency toward overfitting (Mbouopda et al., 2022; Clark et al., 2022) and a lack of generalizability to other locations, as models must be trained separately for each well. Consequently, local models cannot exploit spatial variability or regional dynamics present in different time series across the monitoring network (Bandara et al., 2020).

Global Models

Global models are trained on combined data from multiple monitoring sites and can generate predictions for all locations included in the training set (Mbouopda et al., 2022; Kunz et al., 2024; Heudorfer et al., 2024). This approach enables efficient use of large datasets and facilitates information sharing across the entire network (Clark et al., 2022; Kratzert et al., 2021). Owing to their architecture, global LSTM models function as universal function approximators and can identify and generalize co-occurring patterns across time series (Kratzert et al., 2021). This improves their transferability to ungauged locations or unseen periods (Lees et al., 2022; Feng et al., 2020; Ma et al., 2021). Additional benefits include the ability to model long-memory patterns, robustness to data gaps, and potentially higher computational efficiency compared to training many local models individually (Feng et al., 2020; Chidepudi et al., 2025). In benchmarking studies, global models have frequently outperformed conventional, locally calibrated hydrological models (Kratzert et al., 2019, 2021; Martel et al., 2024; Yu et al., 2024).

However, the general superiority of global models has been questioned in recent studies. A large-scale study by Tran et al. (2025) found that a Google-developed global streamflow model (trained on 5,680 basins) underperformed locally trained single-basin models in 46% of 609 catchments, with peak flows above the 95th–99th percentiles underestimated by an average of -45%. Their meta-analysis of 123 studies further showed that single-basin models frequently achieve excellent skill (NSE ≥ 0.75 in >92% of cases), challenging the view that they are inherently inferior.

These results align with other reported limitations of global models. In the presence of highly heterogeneous time series, performance can decline (Chidepudi et al., 2025; Clark et al., 2022). Global models tend to focus on dominant shared patterns, potentially at the expense of local variability. Learning nonlinear relationships between inputs and targets can become challenging when fundamentally different dynamical behaviors are combined during training (Zhou et al., 2024). Furthermore, studies have shown that static features such as geology, climate, or land use often fail to create proper entity awareness and instead act merely as identifiers (Heudorfer et al., 2024), which limits spatial generalization. Deficits have also been observed for extreme events, for example, due to saturation effects in the LSTM architecture or underestimation of peak values (Baste et al., 2025; Usman et al., 2023; Yu et al., 2024). Finally, the black-box nature of deep neural networks remains a key challenge for decision support in water management, especially for global models, as they capture complex, cross-site patterns that reduce the transparency of local relationships (Gomez et al., 2024; Kratzert et al., 2021; Acuna Espinoza et al., 2024).

Partitioned Models

Partitioned models are essentially global models trained on subgroups of monitoring wells that share similar temporal dynamics or static attributes. These models operate on subgroups of similar time series, which are typically formed through data-driven methods (e.g., clustering algorithms) or domain-specific groupings. The objective is to homogenize training data and to specifically align modeling capacity with similar time series types (Bandara et al., 2020). Clustering is usually based on (i) dynamic time series features such as trend, seasonality, autocorrelation, or entropy (Wunsch et al., 2022b; Gomez et al., 2024); (ii) spectral characteristics to separate typical frequency patterns (Chidepudi et al., 2025); (iii) shape-based similarity metrics such as Grey Relational Clustering (Zhou et al., 2024); or (iv) static catchment attributes like climate, geology, or topography (Kratzert et al., 2024; Kunz et al., 2024). Subsequently, a dedicated model is then trained for each group. Several studies have shown that partitioned models are often more robust to heterogeneity than fully global approaches, particularly when time series exhibit strongly divergent dynamics (Chidepudi et al., 2025). By focusing on homogeneous subgroups, partitioned models can enhance both predictive performance and interpretability (Zhou et al., 2024).

Research Questions and Objectives

In light of these developments, this study aims to systematically compare the predictive performance of global and local deep learning (DL) models for groundwater level forecasting. The central question is whether the advantages of globally trained models, whose superior performance has been widely demonstrated in hydrological streamflow modeling, can be transferred to hydrogeological applications, particularly under the specific conditions of diverse system dynamics; ranging from highly dynamic behavior in karstic aquifers to inertial responses in low-permeability porous aquifers; as well as heterogeneous site conditions and limited data availability.

In contrast to previous studies, the analysis is based on an extensive, Germany-wide groundwater level benchmark dataset comprising nearly 3,000 monitoring wells (Ohmer et al., 2025), spanning over three decades. The associated spatial and dynamic diversity enables a differentiated assessment of the generalizability of data-driven models across different geological and climatic settings.

The core research questions are:

(i) **Overall Model Performance**: Are globally trained LSTM models generally superior to local (single-well) models in terms of overall predictive accuracy across a large and heterogeneous set of monitoring wells?

(ii) **Influence of the Training Data Basis**: How does the predictive performance of global models depend on the characteristics of the training dataset, in particular the number of training wells and the degree of dynamic similarity among them?

(iii) **Prediction of Extreme Events**: Are globally trained models better than single-well models in predicting groundwater-level extremes (e.g., drought lows and high peaks) that were not observed during training? How does predictive performance under extrapolative conditions depend on the size of the training dataset and its degree of dynamic similarity?

(iv) **Out-of-Sample Spatial Prediction**: How well can global models predict groundwater levels at monitoring wells that were not included in the training data?

To address these questions, we conducted a comprehensive experimental comparison. The experiments involve global LSTM models, trained either on the full dataset or on differently partitioned subsets, and locally trained CNN single-well models. All models are evaluated on the same standardized data basis, using test designs that systematically vary the size and dynamic composition of the training dataset, as well as the occurrence of extrapolative conditions such as extreme groundwater levels or unseen locations.

2 Data

2.1 Groundwater Level Data

The analysis is based on the GEMS-GER dataset (Ohmer et al., 2025), which provides standardized groundwater level observations and associated predictor variables for Germany. It contains weekly time series from 3,207 monitoring wells for the period 1991–2022, covering all major hydrogeological regions and a wide range of aquifer types and system dynamics. For this study, we used a filtered subset of 2,951 wells, excluding all sites that achieved an NSE ≤ 0 across all three benchmark models (single-well CNN and global LSTM) described in Ohmer et al. (2025). A detailed description of data sources and quality control procedures is given in the dataset paper. The full dataset is available via Zenodo (DOI: 10.5281/zenodo.15530171). The spatial distribution of monitoring wells is shown in Appendix A (Figure A1), and representative time series in Appendix B (Figure B1).

2.2 Dynamic Input Variables

Each groundwater time series is complemented by dynamic input variables representing meteorological and hydrological conditions, including precipitation, temperature, relative humidity, evapotranspiration, soil moisture, soil temperature, snowmelt, snow water equivalent, and surface as well as subsurface runoff. These variables are provided as part of the GEMS-GER dataset (Ohmer et al., 2025), where they were derived from the HYRAS and ERA5-Land gridded products and preprocessed to a weekly resolution. A detailed description of data sources, derivation methods, and preprocessing steps is given in Ohmer et al. (2025).

2.3 Static Site Attributes

In addition to dynamic inputs, each monitoring well is characterized by a set of more than 50 static attributes, including hydrogeological, topographic, soil, and land use properties. From the full set of static features provided in the dataset (Ohmer et al., 2025), variables related to well depth, screen characteristics, pumping, and pressure state were excluded, as these were sparsely available for the majority of monitoring wells. All categorical static features were label encoded for use in the machine learning models.

3 Methods

3.1 Modeling Strategies

We implemented and compared two main types of modeling strategies: local (single-well) models and global models. The latter were trained either on the full dataset or on differently partitioned subsets (referred to as partitioned models). Throughout this study, we refer to local (single-well) models as **S**, global models as **G**, and partitioned variants as **S-P**x and **G-P**x, where x indicates the partition stage. To indicate the partitioning strategy, the subscript $_{\text{COR}}$ denotes correlation-based removal (increasing dynamic similarity), and the subscript $_{\text{RND}}$ denotes random removal.

(i) Local (Single-Well) Models (S-P0): Independent CNN models were trained for each monitoring well using only local dynamic input variables. These models serve as a site-specific baseline without transferring cross-site information. All single-well models were trained for each of the 2,951 wells (Stage P0). **(ii) Global Model (G-P0):** A single LSTM model was trained on all 2,951 wells of Stage P0 jointly, using both dynamic and static input features to learn generalizable spatio-temporal patterns. **(iii) Partitioned Models (S-P**x**, G-P**x**):** To assess the influence of training set composition, we implemented a series of partitioned models derived from the P0 dataset. The partitioning procedure is defined as follows:

- **Stages P1–P5$_{\text{COR}}$:** Starting from P0, 500 additional wells were successively removed in each stage based on their dynamic dissimilarity to other wells. To quantify this, we computed the pairwise *absolute* Pearson correlations between the standardized groundwater level time series. Each well's dynamic *representativeness* was then defined as the mean absolute correlation with all others. Wells with the lowest representativeness were considered least typical in terms of dynamics and removed first, resulting in subsets with increasing internal similarity.

This strategy was chosen for its transparency, reproducibility, and suitability for systematically reducing dynamic heterogeneity. Compared to more complex, e.g. clustering methods, it avoids hard boundaries and hyperparameter dependencies, offering instead a continuous, interpretable ranking and fine-grained subset control.

- **Stages P1–P5$_{\mathbf{RND}}$:** In parallel, random removal of 500 wells per stage was applied to generate baseline subsets with identical size progression and serve as a control for the correlation-based strategy.

The spatial distribution of monitoring wells and their progressive removal across partitioning stages is shown in Appendix A1.

Global models (G) were retrained on each partitioned subset to reflect the changing training data composition. In contrast, the single-well models (S) were not retrained, as they are inherently independent of other sites. Instead, for each partition stage, only those S models corresponding to the remaining wells were retained for performance evaluation. This approach ensures consistency while enabling a comparative assessment of model robustness under varying training set sizes and internal similarity.

3.2 Model Architectures

All models in this study follow the benchmark architectures introduced in Ohmer et al. (2025), using a standard sequence-to-value forecasting setup. Input sequences of 52 time steps (i.e., weeks) were used to predict the groundwater level at the following time step. Models were trained and validated on the periods 1991–2007 and 2008–2012, respectively, and evaluated on the final 10 years (2013–2022). All metrics were computed from the median prediction of an ensemble of ten independently initialized models.

The **single-well models** are based on a 1D convolutional neural network (CNN) architecture. Each model consists of a convolutional layer with 256 filters and kernel size 3, followed by max pooling, flattening, a dense layer with 32 units, and a final output layer. The models were trained using the Adam optimizer (learning rate 0.001), early stopping (patience 5), a batch size of 16, and a maximum of 30 epochs. Only dynamic input features were used.

The **global models** are based on a Long Short-Term Memory (LSTM) architecture. The dynamic input branch consists of a single LSTM layer with 128 units, followed by a dropout layer with a dropout rate of 0.3. Models were trained for up to 20 epochs using a batch size of 512, early stopping (patience: 5), and a learning rate scheduler targeting a value of 0.001. Static features were incorporated using a second model branch that processes static inputs via a dense layer with 128 units. The outputs of both branches are concatenated and passed through a dense layer with 256 units before the final output layer. Categorical static features were label-encoded. For further architectural and implementation details, we refer to Ohmer et al. (2025).

All global models (G and G-Px) were retrained independently using the same LSTM architecture and hyperparameters, ensuring architectural consistency across all partitioning stages. In contrast, the single-well models (S) were trained once per well on the P0 dataset and remained unchanged; for each partition, only the models corresponding to the retained wells were considered in the evaluation.

3.3 Experimental Design

The experimental design addresses the four research questions outlined in Section 1, each examining a distinct aspect of model performance and generalization. To systematically evaluate these aspects, we conducted four targeted experiments focusing on:

(i) Overall Performance Comparison: We compared the predictive accuracy of global, local, and partitioned models across all monitoring wells (P0). This experiment serves as a baseline to assess overall model performance and consistency across dynamic groundwater regimes.

(ii) Influence of the Training Data Basis: To evaluate how the size and internal similarity of the training dataset affect model performance, we conducted two complementary experiments. First, we compared models trained on subsets with varying degrees of dynamic similarity, created by correlation-based or random well removal (see Section 3.1). Second, we analyzed the effect of progressive random training set size reduction, ranging from 2,951 to 451 wells. This allows us to disentangle the effects of training set size and dynamic consistency on prediction accuracy and robustness.

(iii) Prediction of Extreme Events: Model robustness under extrapolation was assessed by analyzing prediction errors specifically for extreme groundwater conditions, such as droughts or high-water periods. We computed separate metrics for predicted values falling outside the 10th and 90th percentiles of the observed distribution in the test set.

(iv) Out-of-Sample Spatial Prediction: To evaluate the spatial generalization capability of global models, we used the correlation- and random-based partitioning described in Section 3.1. For each stage (P1–P5$_{\mathrm{COR}}$ and P1–P5$_{\mathrm{RND}}$), the excluded wells, i.e., those removed from the training data, served as a spatial out-of-sample test set. This design allows direct assessment of predictive performance at previously unseen locations and isolates the impact of training data composition on generalization.

4 Results and Discussion

The following subsections present the results of the four experiments outlined in Section 3.3, each addressing one of the research questions (RQ i–iv). We evaluate model accuracy and generalization behavior under varying training conditions and hydrological contexts.

4.1 Overall Performance Comparison (RQ i)

To assess whether globally trained LSTM models outperform local single-well (S-P0) models, we compare their predictive performance across 2,951 monitoring wells. Both models were trained on the full dataset (P0).

Figure 1a shows the performance comparison between global (G-P0) and single-well (S-P0) models trained on the full dataset (P0). The upper part displays the distribution of Nash–Sutcliffe efficiency (NSE) values. Overall, the results are fairly similar, with only moderate differences between the global (G-P0) and single-well (S-P0) models. The median NSE is slightly

https://doi.org/10.5194/egusphere-2025-4055
Preprint. Discussion started: 6 October 2025

higher for S-P0 (0.49 vs. 0.47; Table 1), and S-P0 achieves a greater number of high-performing wells, including more values in the upper performance tail. G-P0, in contrast, exhibits a slightly narrower interquartile range, indicating a more compact central distribution. However, the boxplots also reveal that G-P0 produces more wells with very low NSE values, suggesting a higher risk of underperformance at certain locations. However, the boxplots also reveal that G-P0 includes a small number of wells with NSE values lower than those of the corresponding S-P0 models, indicating occasional underperformance at specific locations.

The pairwise comparison in Table 2 confirms these subtle trade-offs: G-P0 outperforms S-P0 at 45.4% of wells, underperforms at 48.9%, and performs equally (within ±0.01 NSE) at 5.7%. In sum, both model types perform broadly comparably, with G-P0 offering slightly more concentrated central performance and S-P0 yielding better results at selected wells. These differences highlight the balance between generalization and local adaptation.

The lower part of Figure 1a shows the cumulative NSE distributions, further illustrating these differences. S-P0 slightly outperforms G-P0 at the very low end (NSE < 0.05), while G-P0 performs better in the range from 0.05 to 0.325. Between 0.325 and 0.425, both models yield nearly identical results. Above this range, S-P0 consistently shows higher cumulative frequencies, indicating better performance in the mid-to-high NSE domain. These findings underline that model performance differences are not uniform across the NSE spectrum. Instead, each model type exhibits advantages in specific performance intervals, without one model consistently outperforming the other across all wells.

4.2 Influence of the Training Data Basis (RQ ii)

To assess how training data characteristics affect global model performance, we analyze two complementary experiments: increasing dynamic similarity through correlation-based well removal, and reducing training data volume (while maintaining the diverse dynamics) through random subsampling.

4.2.1 Dynamic Similarity

To investigate how increasing the internal consistency of the training data affects global model performance, we compare the baseline model G-P0 to a series of partitioned models (G-P1$_{\mathrm{COR}}$ to G-P5$_{\mathrm{COR}}$) trained on increasingly homogeneous subsets. In each step, 500 wells with the lowest average correlation to all other time series were removed to create dynamically more similar training sets.

Figure 1a and Table 1 show that model skill improves consistently with increasing similarity. The share of poorly performing wells (NSE < 0) decreases from 5.5% (G-P0) to 0.0% (G-P5$_{\mathrm{COR}}$), while the proportion of highly accurate wells (NSE > 0.75) increases from 9.8% to 36.6%. The mean NSE rises progressively from 0.47 (G-P0) to 0.54 (G-P1$_{\mathrm{COR}}$), 0.59 (G-P2$_{\mathrm{COR}}$), 0.63 (G-P3$_{\mathrm{COR}}$), 0.65 (G-P4$_{\mathrm{COR}}$), and 0.70 (G-P5$_{\mathrm{COR}}$). The median NSE shows a similar trend, increasing from 0.53 to 0.58, 0.62, 0.66, 0.68, and finally 0.72.

Compared to the corresponding single-well models, global models benefit more strongly from this increased similarity. At stage P5$_{\mathrm{COR}}$, the median NSE of the global model (0.72) exceeds that of the local model (0.67), and the proportion of wells with NSE > 0.75 is nearly twice as high (36.6% vs. 21.1%) (Table 1). Moreover, the share of wells where the global model outperforms its single-well counterpart increases from 45.4% (G-P0) to 67.0% (G-P5$_{\mathrm{COR}}$) (Table 2).

This trend is further illustrated in Figure 2a, which plots global versus local NSE scores across wells for each partition stage. While G-P0 shows many points below the 1:1 line, later stages exhibit a progressive upward shift toward and beyond the diagonal. This indicates that, as training sets become more homogeneous, global models increasingly match or exceed local model performance at individual wells. The point cloud also narrows at higher stages (e.g., P4, P5), reflecting more stable and consistent predictions across sites.

Figure 3a highlights the strong relationship between time series representativeness, quantified as the mean absolute correlation to all other training wells, and model performance. Wells with low representativeness tend to exhibit higher error variance and more frequent underperformance, especially at early stages. From stage P3 onward, a clear threshold emerges around a representativeness value of 0.45, above which consistently high NSE values are achieved. This underscores the central role of dynamic similarity in improving global model skill and reliability.

A qualitative view of these relationships is provided in Figure B1, which displays min–max normalized groundwater level time series for every 20th well, sorted by representativeness, along with the difference in predictive performance (ΔNSE) between single-well and best-performing global models.

Finally, the performance differences between subsequent global models were evaluated on the remaining wells at each stage (Figure 4a). While median ΔNSE values are small, they are consistently positive, and the range of values increases. This shows that the NSE not only increases on average from stage to stage because the wells that are harder to model are removed, but also because the performance of the remaining wells improves compared to the previous stage. In other words, wells with low dynamic representativeness not only pull down the average performance but also hinder the model from achieving higher accuracy on the wells that are more representative.

4.2.2 Training Set Size

To isolate the effect of training data quantity on global model performance, we conducted a second experiment in which wells were randomly removed from the original training set in steps of 500, resulting in five increasingly reduced datasets (G-P1$_{\mathrm{RND}}$ to G-P5$_{\mathrm{RND}}$). In contrast to the correlation-based approach, dynamic similarity was not considered here, allowing us to assess whether model skill improves simply with more training data and whether a critical threshold exists.

Despite the substantial reduction in training data, down to just 451 wells in G-P5$_{\mathrm{RND}}$, global model performance remains remarkably stable. Median NSE values vary only marginally between 0.53 and 0.55, and mean values hover around 0.48 across all stages (Table 1). Similarly, the interquartile range and the overall shape of the NSE distributions (upper part of Figure 1b) show little variation, and the cumulative distribution curves (lower part of Figure 1b) remain largely overlapping. These results suggest that increasing training set size alone does not necessarily lead to better model skill. Interestingly, the global model slightly outperforms the corresponding single-well models in the final stages (P4–P5), reflecting a shift toward more dynamically coherent wells. Thus, while random data reduction does not degrade performance, it also does not yield the benefits commonly associated with larger datasets.

This interpretation is further supported by the summary in Table 2, where the share of poorly performing wells remains around 5%, and the proportion of high-performing wells (NSE > 0.75) increases slightly, from 9.2% to 14.9%, despite the lower number of wells. The global model consistently performs as well as or slightly better than the corresponding local models in the final stages (G > S: 53.2% at $\mathrm{P5}_{\mathrm{RND}}$). While dynamic similarity would be expected to remain constant under random removal, this is not entirely the case for our real-world dataset, as representativeness does not increase continuously but in discrete jumps. At smaller training set sizes, the probability of retaining a more homogeneous subset therefore increases, which can lead to a modest performance gain in later stages. Nevertheless, this improvement is far less pronounced than with correlation-based filtering.

The scatter plots in Figure 2b further support these findings: in contrast to the correlation-based experiment, there is no clear upward shift of the global model scores across stages. Points remain evenly scattered along the 1:1 line, and the share of wells where the global model outperforms the local one increases only marginally. This visual stability reinforces the notion that model skill is mainly independent of training set size, unless accompanied by improved dynamic consistency.

Figure 3b illustrates that, even under random well removal, wells with high representativeness (mean absolute correlation $|\bar{r}| > 0.45$) consistently yield high NSE scores across all stages. However, unlike the correlation-based approach, the representativeness distribution remains broad, and wells with atypical dynamics persist throughout all subsets. As a result, no clear performance threshold emerges and overall skill remains largely stable. There is, however, a slight performance increase of $\mathrm{G}_{\mathrm{RND}}$ across stages, although this effect is modest compared to the gains observed with correlation-based filtering. This small improvement reflects a property of our dataset: representativeness does not increase gradually but in discrete steps, which increases the likelihood that smaller, randomly selected subsets contain more dynamically homogeneous wells. These findings highlight the importance of dynamic similarity rather than dataset size for achieving high predictive skill.

Figure 4b shows that performance differences between consecutive random-reduction stages remain minimal. Median ΔNSE values are consistently close to zero, and the distributions are symmetrically centered, indicating no systematic trend toward improvement or degradation. Although the variance in ΔNSE slightly increases at later stages (e.g., P3–P5), positive and negative deviations remain balanced. This reinforces the interpretation that random data reduction has a neutral net effect on performance: while some wells benefit and others suffer, the overall skill remains unchanged. These findings underline the robustness of the global model to reductions in training data quantity and further support the conclusion that data representativeness is a more critical factor than dataset size.

4.3 Prediction of Extreme Events (RQ iii)

To assess model performance under extrapolative conditions, we evaluated predictions for groundwater levels (GWLs) beyond the typical range observed during training. For each well, low extremes were defined as values in the test period below the 1st percentile of its training distribution, and high extremes as values above the 99th percentile. This site-specific percentile approach ensures that extremes are identified relative to each well's training history, while avoiding dependence on absolute thresholds.

Figure 5 summarizes RMSE distributions for all extrapolated values (top), low extremes (middle), and high extremes (bottom), using both correlation-based and random partitioning. Across all stages, global models do not show improved predictive skill over single-well models. For low extremes, errors are slightly higher for global models at every stage, suggesting that dynamics associated with exceptionally low GWLs are underrepresented in the training sets. For high extremes, both model types perform similarly, with neither showing a consistent advantage.

The stability of error distributions across increasing training set homogeneity or size indicates that, in groundwater systems, larger or more homogeneous datasets do not automatically enhance the prediction of extremes. A plausible explanation is that extreme events often depend on site-specific factors such as fine-scale geology, localized abstraction, or land use, which are not fully captured by the available static site descriptors. Without sufficiently informative descriptors, the transfer of extreme-event knowledge between sites is limited, and events not directly inferable from the dynamic meteorological inputs remain difficult to predict. This reflects a general constraint of current large-scale groundwater datasets rather than a shortcoming of the modeling approach itself.

4.4 Out-of-Sample Spatial Prediction (RQ iv)

To evaluate the spatial transferability of global models, we assessed their performance on monitoring wells deliberately excluded from model calibration. This simulates predictions at sites without prior training data, although observations are available for evaluation. Out-of-sample (OOS) subsets were defined using the partitioning strategies introduced in Section 3.1, i.e., correlation-based removal of dynamically dissimilar wells and random exclusion. Model performance at these OOS sites was compared to that of single-well models trained individually on local data. Figure 6 summarizes the resulting differences in predictive skill.

4.5 Out-of-Sample Spatial Prediction (RQ iv)

To evaluate the spatial transferability of global models, we assessed their performance on monitoring wells deliberately excluded from model calibration. This simulates predictions at sites without prior training data, although observations are available for evaluation. Out-of-sample (OOS) subsets were defined using the partitioning strategies introduced in Section 3.1, i.e., correlation-based removal of dynamically dissimilar wells and random exclusion. Model performance at these OOS sites was compared to that of single-well models trained individually on local data. Figure 6 summarizes the resulting differences in predictive skill.

4.5.1 OOS Based on Dynamic Similarity

The upper panel of Figure 6a summarizes the performance for wells excluded due to dynamic dissimilarity. Global models consistently underperform across all stages, reflecting the difficulty of transferring learned dynamics to sites with low similarity to the training data. In early stages, where excluded wells are most dissimilar, performance deficits are largest. As stages progress, the target wells become more similar to the training set, and global model performance improves, consistent with an increasing likelihood of encountering familiar dynamics. At the same time, however, the number of excluded wells increases, and the training base shrinks, which limits the extent of this improvement.

Despite this trend, the performance gap to single-well models remains largely constant. This suggests that global models, even with broader training data, lack the specificity needed to match the accuracy of locally trained models. Moreover, global predictions show more extreme errors in early stages, indicating that dissimilar wells not only reduce accuracy but also increase the risk of model failure. Overall, these results underline the limited spatial generalization capacity of global models under strong dynamic heterogeneity.

The lower panel of Figure 6a shows the cumulative distribution of NSE values for the out-of-sample wells. Across all stages, the curves for the global models lie consistently below those of the single-well models, confirming their overall weaker performance. While the global distributions shift slightly rightward with increasing stage, reflecting improved prediction accuracy as the excluded wells become more similar to the training set, the performance gap remains substantial. Notably, the curves diverge across nearly the entire NSE range, indicating that the deficit is not limited to a few poor predictions but affects a broad range of wells. This reinforces the limited ability of global models to generalize to dynamically dissimilar systems.

4.5.2 OSS Random Based

In the random partitioning, wells are excluded from training regardless of dynamic similarity. This ensures a structurally balanced training set across stages, while the number of OOS predictions increases from 500 to 2,500 (and the number of training wells decreases from 2,451 to 451). Figure 6b summarizes the results.

The upper part of Figure 6b shows the NSE differences between global and single-well models across stages. Global models perform slightly worse throughout, with median differences becoming gradually more negative as the model is required to predict an increasing number of wells it has never seen during training. While the performance gap remains small, it reflects the growing difficulty of maintaining generalization under data reduction. The distribution width remains comparable to that of the single-well models, and extreme outliers are rare, suggesting that random data removal does not cause systematic prediction failures.

The lower part of Figure 6b displays the cumulative distribution of NSE values for all OOS wells. Global and single-well curves are closely aligned across all stages, confirming that both approaches yield broadly similar predictive skill under random exclusion. A slight tendency toward underperformance remains visible for the global models, particularly in the mid-to-upper NSE range. These results emphasize that while global models generalize reasonably well in structurally balanced settings, they do not gain measurable advantage from larger training sets and remain slightly inferior to locally specialized models, even under idealized conditions of random exclusion.

5 Discussion

5.1 Comparison with previous studies

Our findings provide mixed support for earlier results from deep learning applications in hydrology and hydrogeology. In line with studies in streamflow modeling (Kratzert et al., 2019, 2024; Martel et al., 2024), we find that global models can achieve predictive skill comparable to or exceeding that of local models when trained on large dynamically homogeneous datasets. This confirms the general advantage of cross-site learning in environments where system dynamics are similar, as also observed in other partitioning approaches (Chidepudi et al., 2025; Zhou et al., 2024; Clark et al., 2022).

However, while global models have often shown clear advantages for streamflow applications (Kratzert et al., 2021; Lees et al., 2022), though not without recent dissenting findings (Tran et al., 2025), our results for groundwater level prediction reveal no overall advantage under heterogeneous conditions.

This difference is likely related to the broader diversity and high small-scale variability of groundwater system dynamics, ranging from highly responsive karst aquifers to inertial systems in low-permeability sediments. Unlike many surface water catchments, groundwater dynamics can differ markedly even over short distances. Nearby wells may share similar static feature values (e.g., geology, land use) yet exhibit distinct responses due to fine-scale geological differences, flow paths, or localized abstraction.

Under such conditions, a global model may still learn the overall diversity of dynamics present in the training set but, due to insufficiently informative static site descriptors, lacks the ability to reliably assign the correct dynamic behavior to a specific well. As a result, the model tends to average across similar but not identical dynamics, which can reduce accuracy at individual sites. This mechanism could explain why global models in our experiments did not achieve the same consistent advantage as reported in streamflow studies when trained on heterogeneous data.

Consistent with the concerns raised by Heudorfer et al. (2024), the static site descriptors available in our dataset (geology, climate, land use) are not sufficiently informative to enable strong entity awareness. This limits the model's ability to transfer knowledge in out-of-sample predictions and in forecasting extreme events absent from a site's training period, often resulting in more generic, averaged outputs. The descriptors used here represent the best practical dataset currently available for large-scale groundwater applications, so this limitation reflects a general constraint of current large-scale groundwater modeling rather than a shortcoming of the modeling approach itself. Such limitations may explain why, contrary to several streamflow studies (Baste et al., 2025; Yu et al., 2024; Kratzert et al., 2024), neither global nor local models in our experiments showed a substantial advantage in predicting extreme events outside the training range.

Finally, our partitioning experiments confirm the robustness benefits reported in other hydrological contexts (Chidepudi et al., 2025; Zhou et al., 2024): grouping wells with similar dynamics before training significantly improved the performance of global models, even with fewer training wells. This supports the broader conclusion from the literature that data homogeneity, whether achieved via targeted filtering or clustering, can be more important for generalization skill than sheer dataset size. This further reinforces that, in groundwater modeling, dynamic similarity often outweighs data quantity as a determinant of global model skill.

6 Conclusion

This study provides a comprehensive evaluation of globally and locally trained deep learning models for groundwater level forecasting. Using a dataset comprising nearly 3,000 monitoring wells across Germany, we systematically assessed model performance across diverse hydrogeological settings and under varying conditions of data availability, dynamic similarity, and extrapolation demands. The analysis was guided by four research questions, each addressing a key aspect of model generalization and applicability. Below, we summarize the main findings in response to each question, placing them in the context of previous hydrological research.

(i) Are globally trained models generally superior to local (single-well) models?

Not necessarily. Despite being trained on a large and diverse dataset, globally trained models did not show a notable overall advantage over locally optimized single-well models. Local models tend to achieve slightly higher median accuracy and perform better at individual sites, while global models produce more predictions clustered around the central range of performance, suggesting a more robust but less specialized behavior. These findings align with earlier work in groundwater modeling but contrast with the consistent superiority reported for streamflow, likely reflecting the greater diversity and small-scale variability of groundwater system dynamics.

(ii) Does training data quality or quantity matter more for global models?

Training data quality in terms of dynamic similarity is more important than quantity. When the training set is filtered to include only wells with similar temporal dynamics, global model accuracy improves markedly. In contrast, random removal of wells (reducing size without regard to similarity) does not improve performance, even when up to 85% of the data are removed. Quantity without structure provides no measurable benefit. For in-sample predictions, dynamic similarity is clearly the dominant factor; for out-of-sample predictions, a broader and more diverse training base can, in theory, offer slight robustness gains, although in our experiments this effect was minimal.

(iii) Can global models reliably predict extreme groundwater events?

No. In our experiments, both global and local models consistently failed to provide accurate predictions for groundwater levels outside the training range. Across all partitioning stages and extrapolation regimes, global models showed slightly higher errors and a tendency to overestimate low values while underestimating high ones, reflecting a structural averaging effect. Neither increasing the amount of training data nor improving dynamic similarity mitigated this issue. A likely reason is that the available static site descriptors (e.g., geology, land use, geomorphology) are not sufficiently informative to provide strong entity awareness, thereby limiting the transfer of extreme-event knowledge between sites. These descriptors represent the best practical dataset currently available for large-scale groundwater modeling, so this limitation reflects a general constraint of current data availability rather than a shortcoming of the modeling approach itself.

(iv) How well do global models generalize to unseen locations?

The global model shows limited spatial generalization, strongly depending on the similarity between training and target wells. Under correlation-based exclusion, performance drops sharply compared to local models, reflecting the difficulty of transferring learned dynamics to dissimilar sites. Even as similarity increases with successive stages, the gap to single-well models remains substantial. In contrast, random exclusion yields broadly similar performance, indicating that generalization is feasible when target wells share representative temporal patterns with the training data, consistent with previous findings that groundwater dynamics are less transferable than streamflow.

In summary, the choice between global and local modeling depends strongly on the intended application. For in-sample prediction at known sites, training on dynamically similar wells (e.g., by partitioning the dataset through clustering) can yield accurate results even with relatively few data, as the remaining wells benefit from the removal of poorly representative sites. For spatial generalization, a broad and diverse training base may increase robustness, though often at the cost of predictive precision. For groundwater systems, characterized by slow and often indirect responses, sparse measurements, high small-scale variability, and limited entity awareness due to coarse static descriptors, model transferability appears inherently more limited than in surface water applications. This highlights that single-well models remain a strong option, especially when site-specific accuracy is required and regional dynamics need to be captured in detail.

Code and data availability. The code used in this study is available on (https://github.com/KITHydrogeology/singlewell-vs-global-gwl), and the data are provided via Zenodo (https://doi.org/10.5281/zenodo.15530171).

Table 1. Overview of performance metrics for global (G) and corresponding single-well (S) models, including Nash–Sutcliffe efficiency (NSE), root mean square error (RMSE), coefficient of determination (R^2), and bias. Values are reported as minimum, median, mean, and maximum across all wells for each model configuration.

Model	**NSE**				**RMSE**				**R^2**				**Bias**			
	min	med	mean	max	min	med	mean	max	min	med	mean	max	min	med	mean	max
G-P0	-1.16	0.53	0.47	0.91	0.02	0.25	0.37	8.90	0.00	0.62	0.57	0.93	-7.89	0.02	0.04	6.33
G-P1$_{\text{COR}}$	-1.22	0.58	0.54	0.91	0.04	0.24	0.33	7.02	0.00	0.65	0.62	0.93	-4.86	0.03	0.04	4.59
G-P2$_{\text{COR}}$	-0.71	0.62	0.59	0.92	0.04	0.23	0.32	7.39	0.08	0.69	0.66	0.94	-1.59	0.02	0.04	4.36
G-P3$_{\text{COR}}$	-0.51	0.66	0.63	0.93	0.04	0.22	0.30	7.35	0.08	0.71	0.69	0.94	-1.25	0.03	0.05	3.54
G-P4$_{\text{COR}}$	-0.51	0.68	0.65	0.92	0.05	0.22	0.29	7.23	0.13	0.73	0.71	0.94	-0.64	0.03	0.04	3.52
G-P5$_{\text{COR}}$	0.20	0.72	0.70	0.94	0.05	0.20	0.26	5.05	0.36	0.77	0.75	0.96	-0.50	0.03	0.05	3.58
G-P1$_{\text{RND}}$	-1.27	0.54	0.47	0.90	0.02	0.25	0.38	9.01	0.00	0.62	0.57	0.93	-7.92	0.02	0.05	6.53
G-P2$_{\text{RND}}$	-1.08	0.54	0.48	0.91	0.02	0.25	0.38	7.85	0.00	0.63	0.58	0.93	-7.80	0.03	0.06	6.43
G-P3$_{\text{RND}}$	-0.96	0.55	0.48	0.89	0.02	0.25	0.37	7.69	0.00	0.63	0.58	0.94	-5.62	0.03	0.06	6.25
G-P4$_{\text{RND}}$	-1.07	0.55	0.49	0.88	0.02	0.24	0.36	8.00	0.00	0.63	0.58	0.92	-5.26	0.03	0.07	6.59
G-P5$_{\text{RND}}$	-1.10	0.57	0.50	0.90	0.04	0.24	0.34	5.83	0.00	0.62	0.58	0.93	-5.30	0.03	0.05	5.40
S-P0	-1.21	0.55	0.49	0.94	0.02	0.25	0.36	8.06	0.00	0.59	0.54	0.93	-5.78	0.03	0.07	6.34
S-P1$_{\text{COR}}$	-0.53	0.59	0.54	0.94	0.04	0.24	0.33	8.06	0.00	0.63	0.59	0.93	-2.25	0.03	0.06	5.63
S-P2$_{\text{COR}}$	-0.53	0.62	0.57	0.94	0.04	0.23	0.33	8.06	0.00	0.66	0.62	0.93	-2.25	0.03	0.06	5.63
S-P3$_{\text{COR}}$	-0.41	0.64	0.60	0.94	0.04	0.23	0.31	8.06	0.01	0.67	0.64	0.93	-1.13	0.04	0.06	5.63
S-P4$_{\text{COR}}$	-0.41	0.65	0.62	0.94	0.06	0.22	0.31	8.06	0.13	0.68	0.66	0.93	-0.46	0.04	0.06	5.63
S-P5$_{\text{COR}}$	-0.27	0.67	0.64	0.89	0.07	0.21	0.29	7.16	0.24	0.69	0.67	0.92	-0.46	0.05	0.07	5.63
S-P1$_{\text{RND}}$	-1.21	0.55	0.49	0.94	0.02	0.25	0.37	8.06	0.00	0.59	0.54	0.93	-5.78	0.03	0.07	6.34
S-P2$_{\text{RND}}$	-1.21	0.55	0.48	0.94	0.02	0.25	0.37	8.06	0.00	0.59	0.54	0.92	-5.78	0.03	0.07	6.34
S-P3$_{\text{RND}}$	-0.83	0.55	0.48	0.94	0.02	0.25	0.36	8.06	0.00	0.59	0.54	0.92	-4.09	0.03	0.07	6.34
S-P4$_{\text{RND}}$	-0.83	0.54	0.48	0.91	0.02	0.25	0.36	8.06	0.00	0.58	0.54	0.92	-3.68	0.04	0.07	6.34
S-P5$_{\text{RND}}$	-0.83	0.53	0.48	0.90	0.04	0.25	0.34	5.99	0.00	0.57	0.53	0.92	-3.68	0.04	0.06	4.97

Table 2. Summary of model performance across correlation- and randomly reduced training subsets. Left columns show NSE-based performance groups for global models (G), middle columns the corresponding results for local single-well models (S) trained on the same well subsets. Right columns report the share of wells for which global models perform better, worse, or equally (±0.01 NSE) compared to their local counterparts.

G-Model	**<0**	**>0.65**	**>0.75**	**>0.85**	**S-Model**	**<0**	**>0.65**	**>0.75**	**>0.85**	**G >**	**S >**	**Equal**
G-P0	5.5	27.4	9.8	1.3	S-P0	5.0	32.2	13.1	2.5	45.4	48.9	5.7
G-P1$_{\text{COR}}$	2.1	34.4	11.4	1.3	S-P1$_{\text{COR}}$	2.8	37.0	15.3	2.9	47.2	45.9	6.9
G-P2$_{\text{COR}}$	0.9	44.2	17.9	2.1	S-P2$_{\text{COR}}$	2.0	42.1	17.6	3.4	50.0	43.4	6.6
G-P3$_{\text{COR}}$	0.3	52.2	24.2	3.0	S-P3$_{\text{COR}}$	1.3	46.8	19.7	3.9	53.5	38.9	7.6
G-P4$_{\text{COR}}$	0.1	58.0	27.5	2.7	S-P4$_{\text{COR}}$	0.9	51.5	20.5	3.4	56.5	36.9	6.6
G-P5$_{\text{COR}}$	0.0	71.8	36.6	6.9	S-P5$_{\text{COR}}$	0.7	57.6	21.1	3.8	67.0	27.3	5.8
G-P1$_{\text{RND}}$	5.3	26.5	8.9	1.0	S-P1$_{\text{RND}}$	5.0	32.1	12.9	2.4	44.4	50.1	5.4
G-P2$_{\text{RND}}$	5.1	29.0	11.2	1.6	S-P2$_{\text{RND}}$	4.9	31.7	12.5	2.4	48.8	45.4	5.8
G-P3$_{\text{RND}}$	5.1	30.2	11.2	1.4	S-P3$_{\text{RND}}$	4.9	31.2	12.9	2.5	46.9	46.3	6.8
G-P4$_{\text{RND}}$	4.5	31.3	11.5	1.2	S-P4$_{\text{RND}}$	5.0	30.6	12.5	2.2	48.9	44.2	6.9
G-P5$_{\text{RND}}$	5.1	32.8	14.9	1.6	S-P5$_{\text{RND}}$	4.2	29.3	12.2	2.2	53.2	39.5	7.3

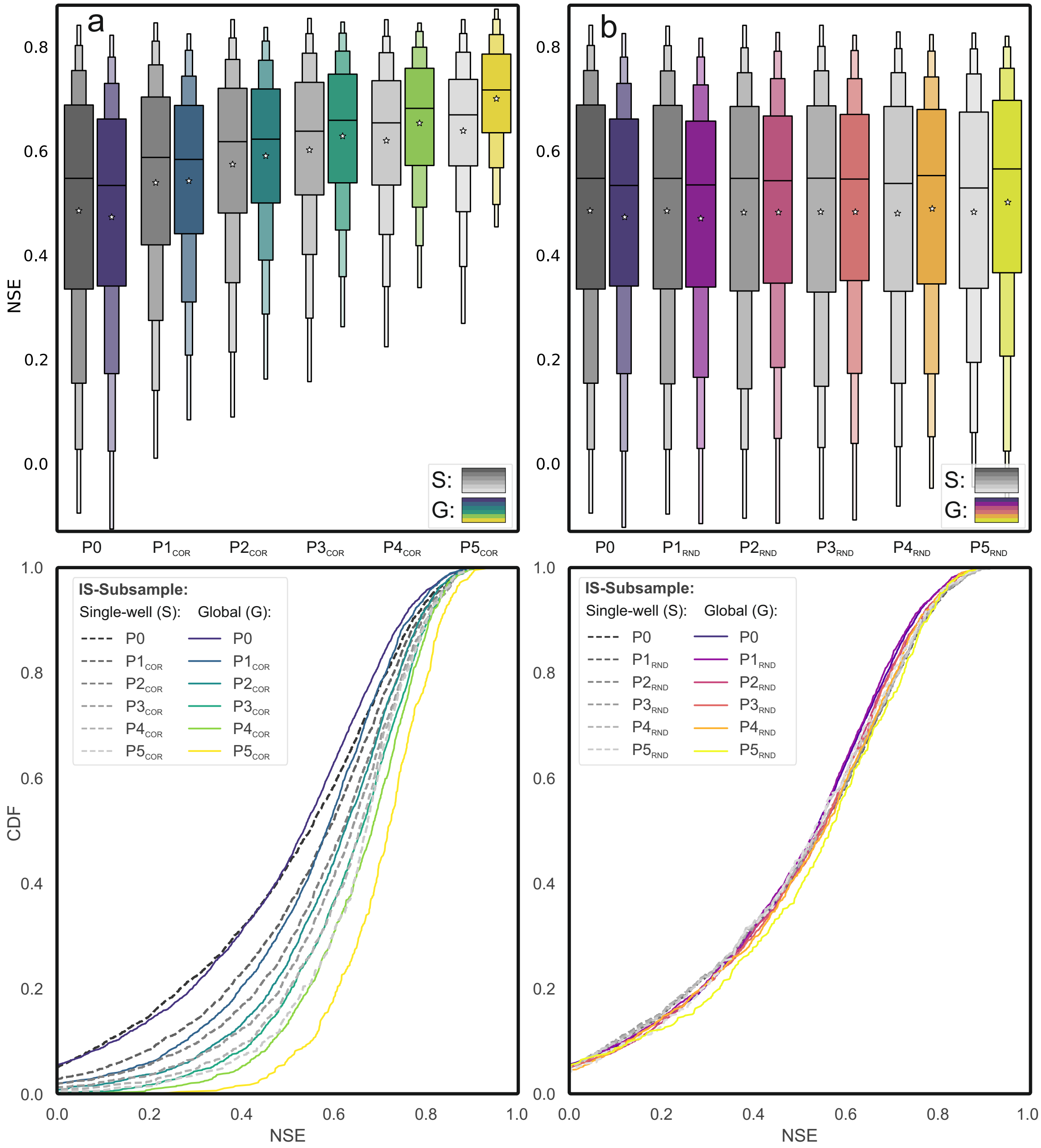

Figure 1. Comparison of single-well and global model performance across generalization stages. (a, b) Distributions of NSE scores (*top*) and cumulative distribution functions (*bottom*) for single-well (S) and global models (GP0–GP5), based on either correlation-based (a, GP_{cor}) or random (b, GP_{rnd}) well selection. Each global model is compared to S models trained on the same subset, illustrating shifts in performance distributions with increasing training data homogeneity (a) or quantity (b).

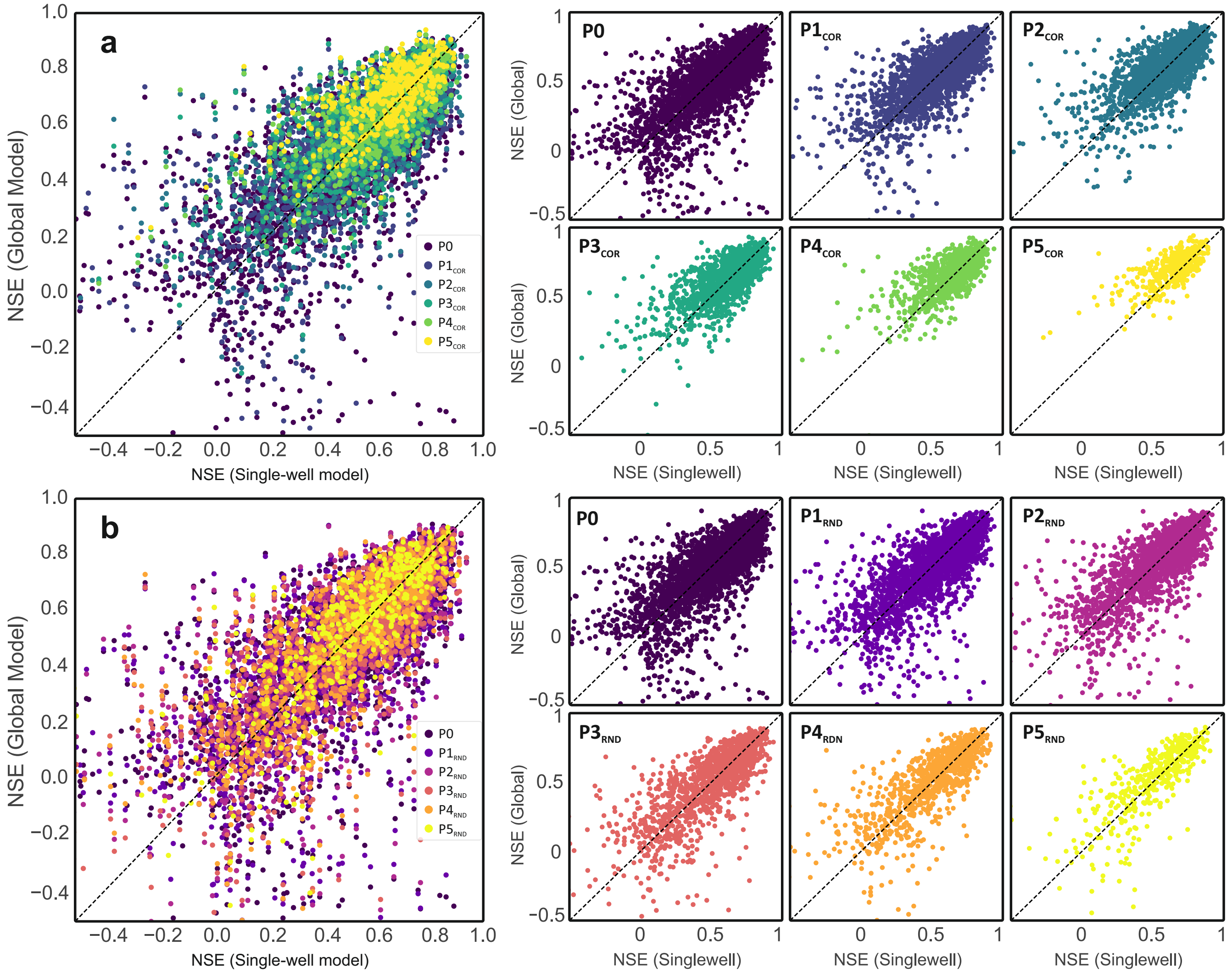

Figure 2. Comparison of global and single-well model performance at the well level. Panels (a) and (b) show NSE values of global models (G-P_x) plotted against their corresponding single-well models (S-P_x) for each monitoring well. Panel (a) includes models trained on dynamically similar subsets (G-P_{COR}), and panel (b) shows models trained on randomly selected subsets (G-P_{RND}). Colored points indicate generalization stages (P0–P5). Right-hand subplots display the same data disaggregated by stage. Points above the 1:1 line mark wells where the global model outperforms its local counterpart.

https://doi.org/10.5194/egusphere-2025-4055
Preprint. Discussion started: 6 October 2025

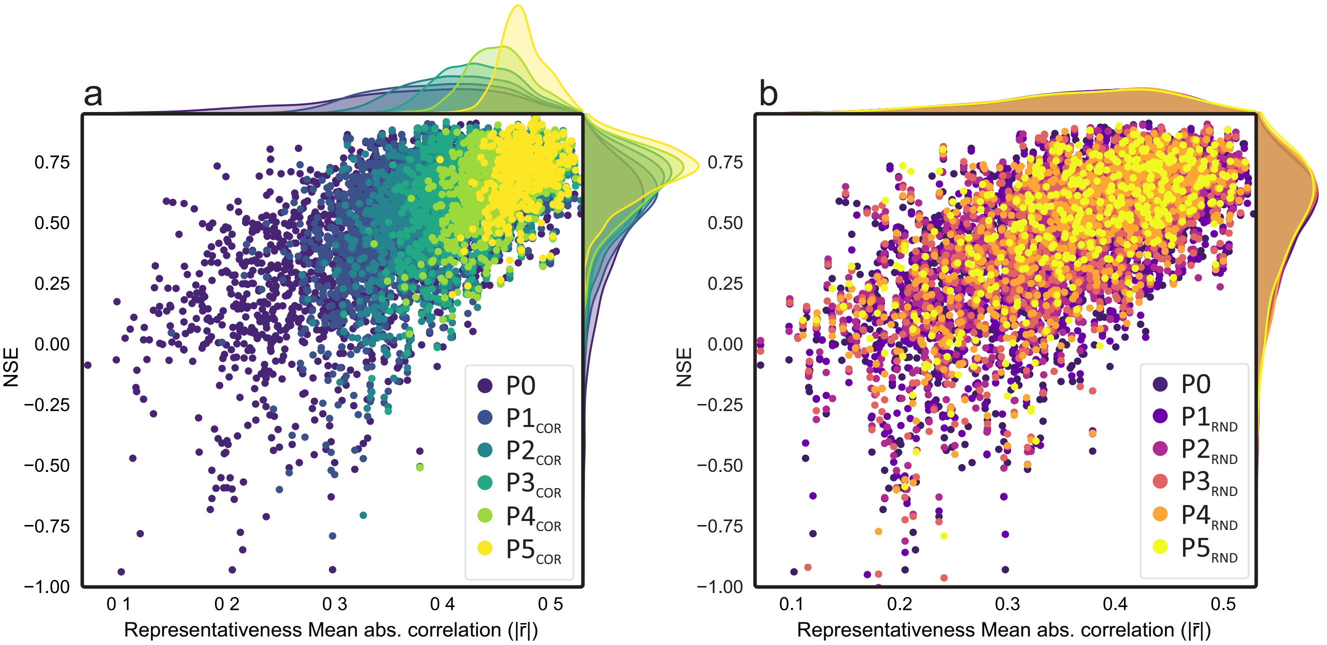

Figure 3. Relationship between time series representativeness and model performance. NSE scores of global models are plotted against the representativeness of each well, defined as the mean absolute correlation with all other training wells. Panel (a) shows results for correlation-based removal (P1–P5$_{\mathrm{COR}}$), and panel (b) for random removal (P1–P5$_{\mathrm{RND}}$). Densities along the top axis indicate the distribution of representativeness across generalization stages (P0–P5). Model performance increases with higher representativeness, particularly under the COR setting, where wells with atypical dynamics are systematically excluded.

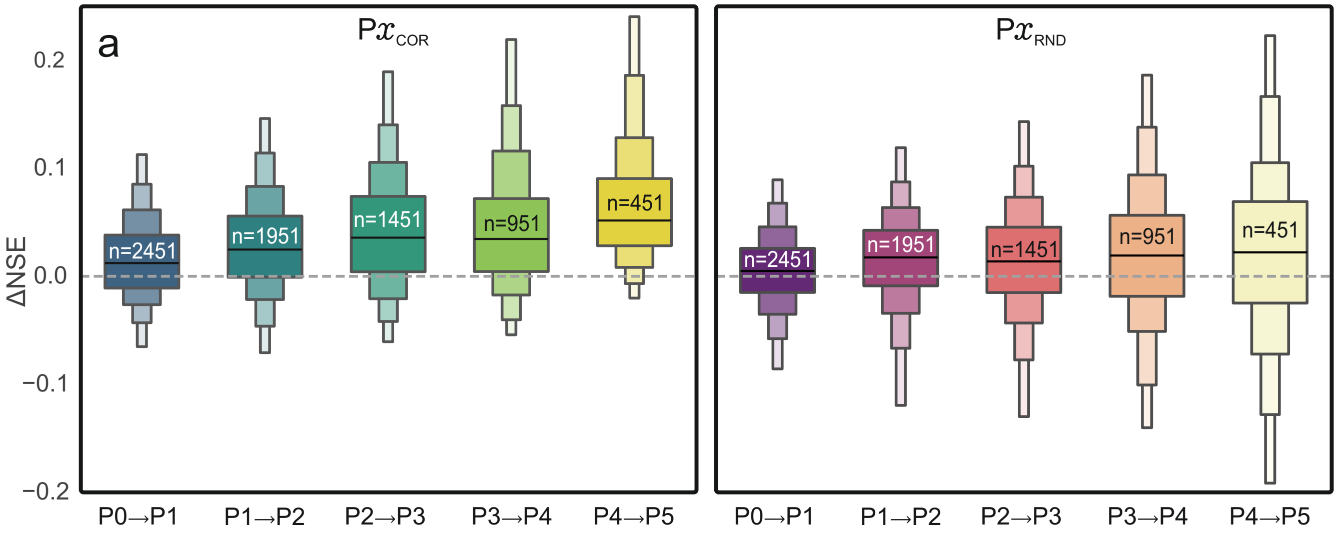

Figure 4. Change in model performance across generalization stages. Distributions of ΔNSE (difference in NSE) between successive global models trained on progressively smaller training sets. Panel (a) shows correlation-based stages (GP_{COR}), and panel (b) random stages (GP_{RND}). Each comparison quantifies the performance difference between two consecutive stages (e.g., GP3 vs. GP2) for the remaining training wells. Boxes indicate the number of wells included in each comparison. Median ΔNSE values remain close to zero, suggesting no systematic loss in predictive accuracy with increasing data reduction.

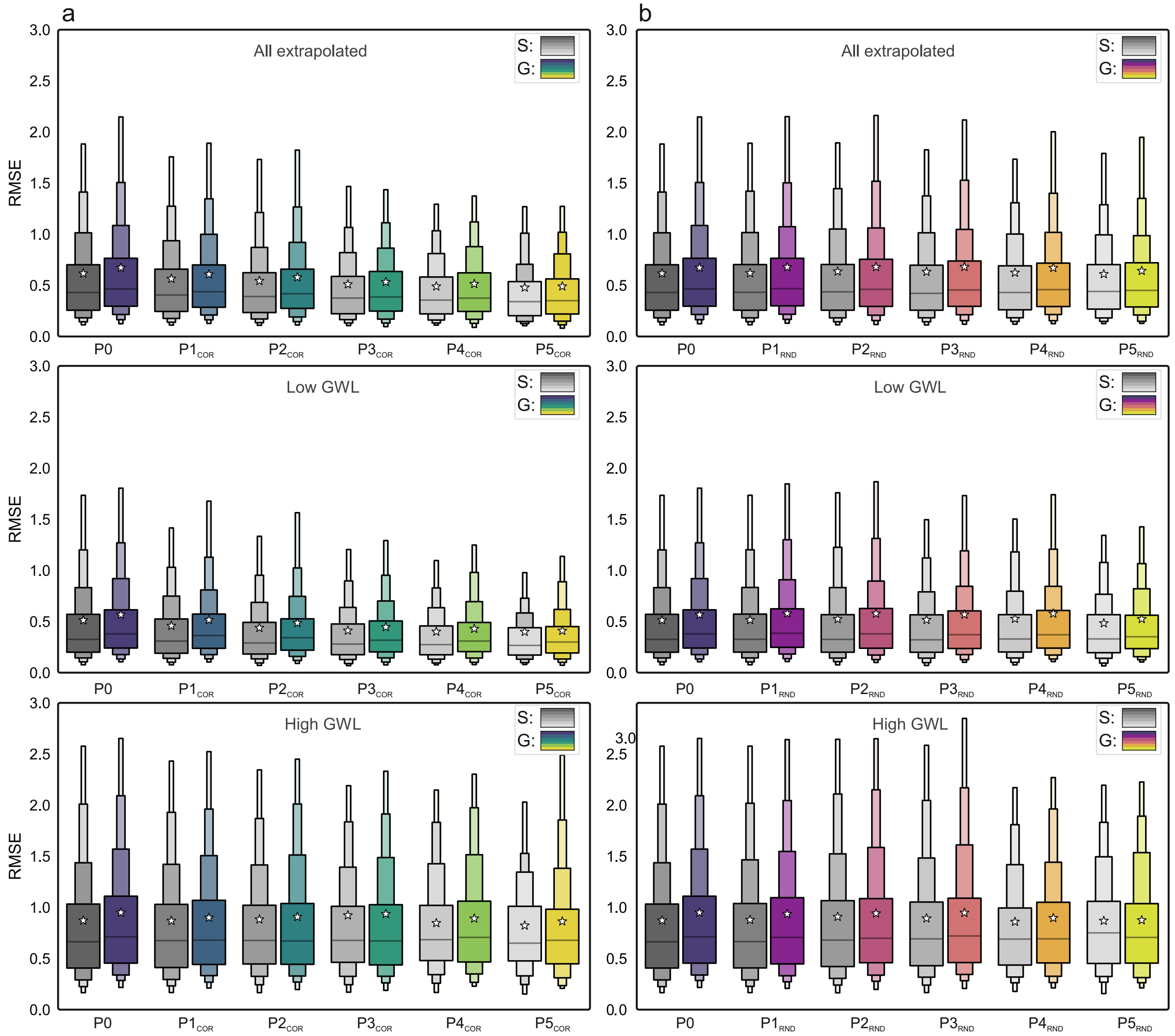

Figure 5. Model performance under extrapolated conditions. Boxplots of RMSE for single-well (S) and global (G) models across generalization stages (P0–P5). Panel (a) shows correlation-based stages ($\mathrm{GP_{COR}}$), and panel (b) random stages ($\mathrm{GP_{RND}}$). Each panel displays results for all extrapolated time steps (*top*), as well as separated by low groundwater levels (*middle*) and high groundwater levels (*bottom*).

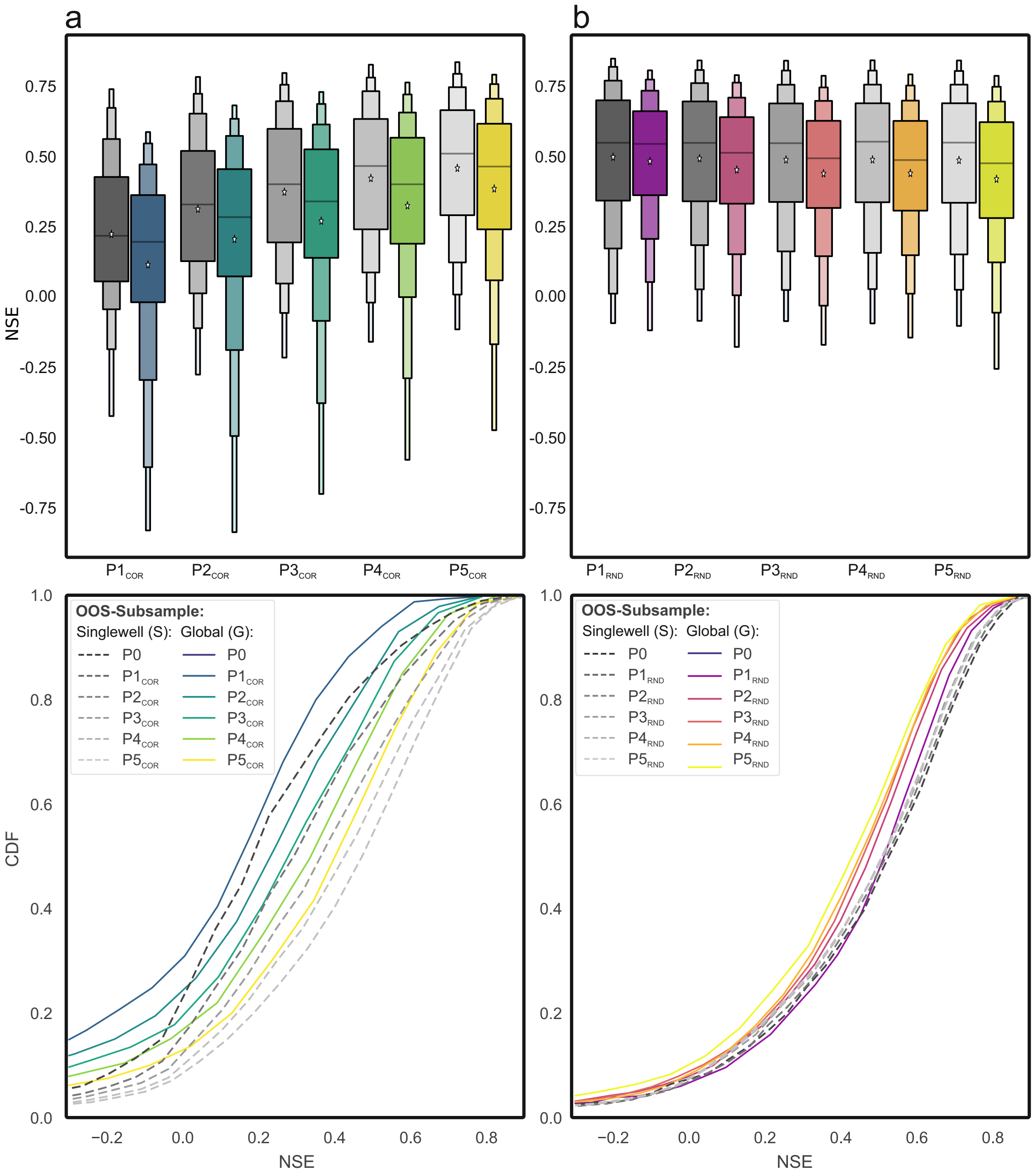

Figure 6. Comparison of single-well and global model performance across generalization stages (out-of-sample wells). Boxplots (*top*) and cumulative distribution functions (CDFs, *bottom*) of NSE for single-well (S) and global (G) models across stages P1–P5. Panel (a) shows correlation-based stages (GP_{COR}), and panel (b) random stages (GP_{RND}). Each global model is compared to S models trained on the same subset of wells, illustrating performance differences under spatial extrapolation.

https://doi.org/10.5194/egusphere-2025-4055
Preprint. Discussion started: 6 October 2025

450 Appendix A: Spatial distribution of monitoring wells

Figure A1 shows the geographical distribution of all monitoring wells used in the modeling experiments, as well as their progressive removal across different partitioning stages.

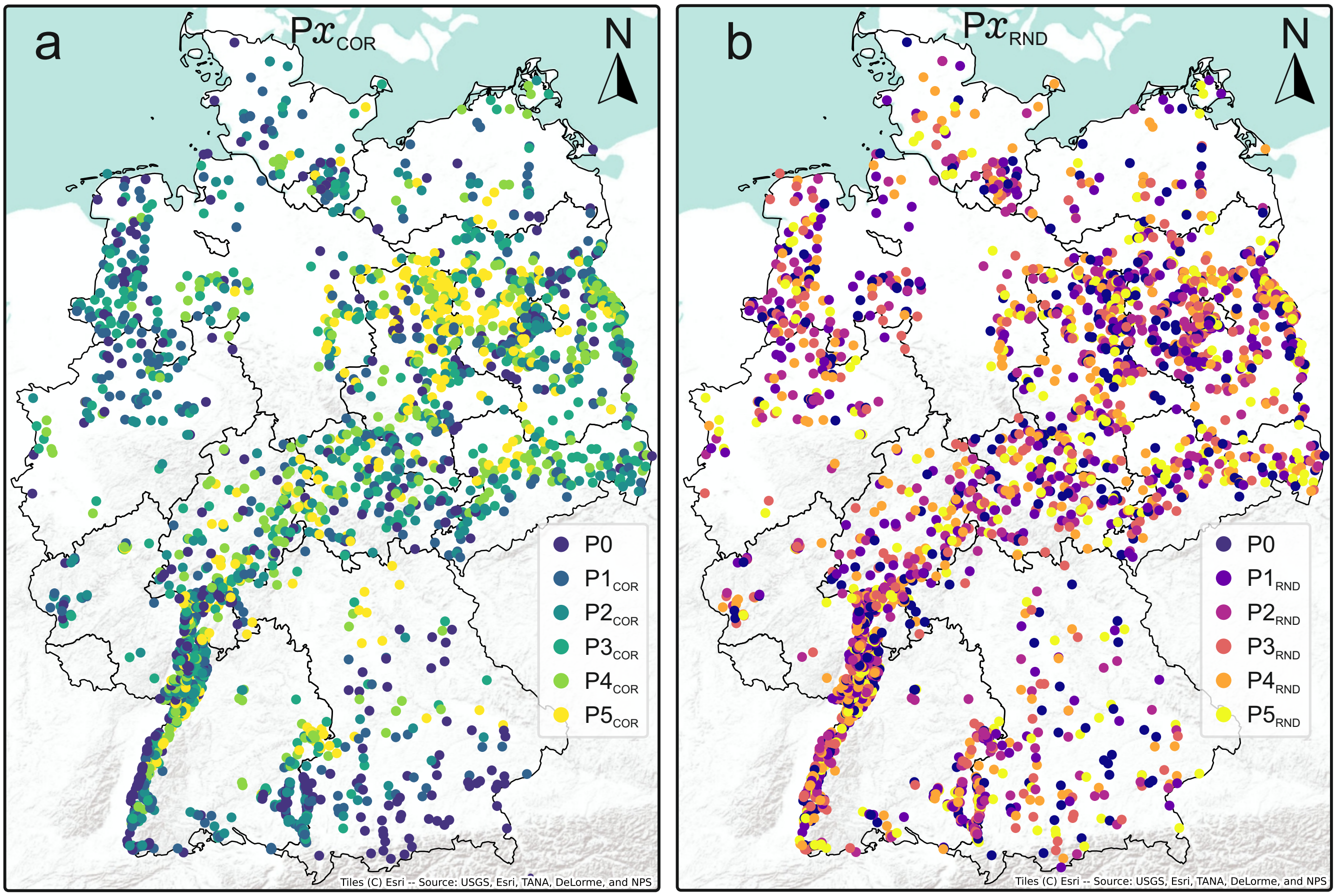

Figure A1. Spatial distribution of groundwater monitoring wells used in this study. The panels distinguish between correlation-based (P_{COR}, a) and random (P_{RND}, b) data removal scenarios across six stages (P0–P5). Each stage represents a progressive reduction of the training data set, either by removing wells with low dynamic similarity (P_{COR}) or through random subsampling (P_{RND}). The map highlights how spatial coverage changes with increasing data reduction

https://doi.org/10.5194/egusphere-2025-4055
Preprint. Discussion started: 6 October 2025

Appendix B: Stacked groundwater level time series with representativeness and performance difference

Figure B1 shows min–max normalized groundwater level time series for every 20th monitoring well (from the second-highest 455 representativeness rank), ordered by representativeness.

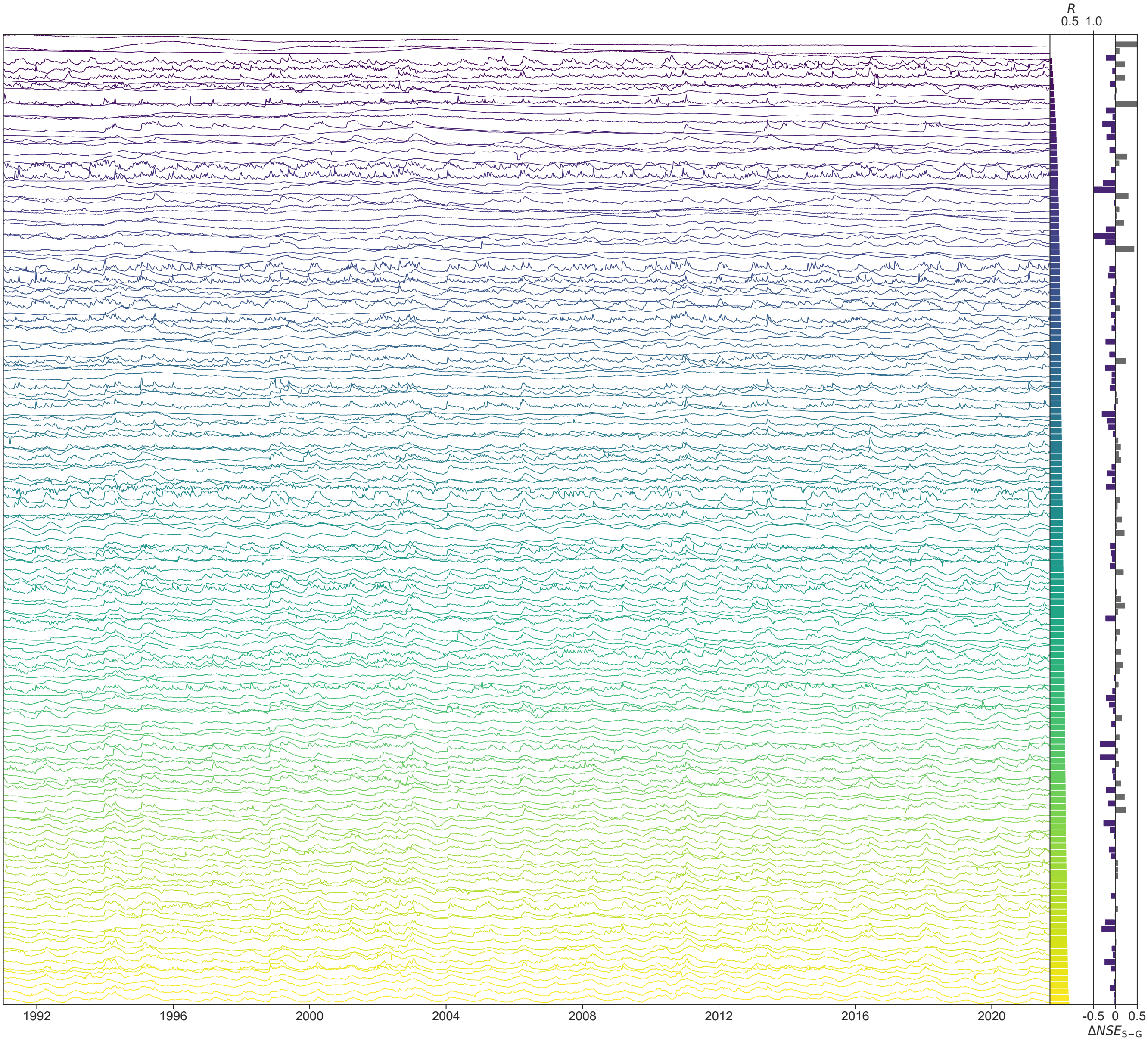

Figure B1. Stacked groundwater level (GWL) time series (min–max normalized) by representativeness. Left: time series color-coded by representativeness (R); middle: bars of R; right: ΔNSE = NSE$_S$ − NSE$_G$, clipped to $[-0.5, 0.5]$ (dark gray = S better, blue = G better).

Author contributions. MO carried out the data analysis, prepared the figures and plots, and drafted the main part of the manuscript. Conceptualisation and methodology were developed jointly by MO and TL. TL conducted most of the modelling experiments and contributed to the interpretation of results. Both authors revised and approved the final manuscript.

Competing interests. The authors declare that they have no conflict of interest

460 *Acknowledgements.* All programming was done in Python version 3.12 (van Rossum, 1995) and the associated libraries, including NumPy (Harris et al., 2020), Pandas (McKinney, 2010), Tensorflow (Abadi et al., 2016), Keras (Chollet, 2015), SciPy (Virtanen et al., 2020), Scikit-learn (Pedregosa et al., 2011) and Matplotlib (Hunter, 2007). The authors further acknowledge support by the state of Baden-Württemberg through bwHPC.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X.: TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, in: 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), https://arxiv.org/abs/1603.04467, arXiv:1603.04467, 2016.

Acuna Espinoza, E., Loritz, R., Kratzert, F., Klotz, D., Gauch, M., Álvarez Chaves, M., Bäuerle, N., and Ehret, U.: Analyzing the generalization capabilities of hybrid hydrological models for extrapolation to extreme events, https://doi.org/10.5194/egusphere-2024-2147, 2024.

Bandara, K., Bergmeir, C., and Smyl, S.: Forecasting across time series databases using recurrent neural networks on groups of similar series: A clustering approach, Expert Systems with Applications, 140, 112 896, https://doi.org/10.1016/j.eswa.2019.112896, publisher: Elsevier BV, 2020.

Baste, S., Klotz, D., Espinoza, E. A., Bardossy, A., and Loritz, R.: Unveiling the Limits of Deep Learning Models in Hydrological Extrapolation Tasks, https://doi.org/10.5194/egusphere-2025-425, 2025.

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?, Hydrology and Earth System Sciences, 29, 841–861, https://doi.org/10.5194/hess-29-841-2025, publisher: Copernicus GmbH, 2025.

Chollet, F.: Keras, https://github.com/fchollet/keras, 2015.

Clark, S. R., Pagendam, D., and Ryan, L.: Forecasting Multiple Groundwater Time Series with Local and Global Deep Learning Networks, International Journal of Environmental Research and Public Health, 19, 5091, https://doi.org/10.3390/ijerph19095091, publisher: MDPI AG, 2022.

Feng, D., Fang, K., and Shen, C.: Enhancing Streamflow Forecast and Extracting Insights Using Long-Short Term Memory Networks With Data Integration at Continental Scales, Water Resources Research, 56, https://doi.org/10.1029/2019wr026793, publisher: American Geophysical Union (AGU), 2020.

Gomez, M., Nölscher, M., Hartmann, A., and Broda, S.: 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, 4407–4425, https://doi.org/10.5194/hess-28-4407-2024, publisher: Copernicus GmbH, 2024.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E.: Array programming with NumPy, Nature, 585, 357–362, https://doi.org/10.1038/s41586-020-2649-2, 2020.

Hauswirth, S. M., Bierkens, M. F., Beijk, V., and Wanders, N.: The potential of data driven approaches for quantifying hydrological extremes, Advances in Water Resources, 155, 104 017, https://doi.org/10.1016/j.advwatres.2021.104017, 2021.

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

Hunter, J. D.: Matplotlib: A 2D Graphics Environment, Computing in Science & Engineering, 9, 90–95, https://doi.org/10.1109/MCSE.2007.55, 2007.

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, Hydrology and Earth System Sciences, 23, 5089–5110, https://doi.org/10.5194/hess-23-5089-2019, publisher: Copernicus GmbH, 2019.

Kratzert, F., Gauch, M., Nearing, G., Hochreiter, S., and Klotz, D.: Niederschlags-Abfluss-Modellierung mit Long Short-Term Memory (LSTM), Österreichische Wasser- und Abfallwirtschaft, 73, 270–280, https://doi.org/10.1007/s00506-021-00767-z, publisher: Springer Science and Business Media LLC, 2021.

Kratzert, F., Gauch, M., Klotz, D., and Nearing, G.: HESS Opinions: Never train a Long Short-Term Memory (LSTM) network on a single basin, Hydrology and Earth System Sciences, 28, 4187–4201, https://doi.org/10.5194/hess-28-4187-2024, publisher: Copernicus GmbH, 2024.

Kunz, S., Schulz, A., Wetzel, M., Nölscher, M., Chiaburu, T., Biessmann, F., and Broda, S.: Towards a Global Spatial Machine Learning Model for Seasonal Groundwater Level Predictions in Germany, https://doi.org/10.5194/egusphere-2024-3484, 2024.

Lees, T., Reece, S., Kratzert, F., Klotz, D., Gauch, M., De Bruijn, J., Kumar Sahu, R., Greve, P., Slater, L., and Dadson, S. J.: Hydrological concept formation inside long short-term memory (LSTM) networks, Hydrology and Earth System Sciences, 26, 3079–3101, https://doi.org/10.5194/hess-26-3079-2022, publisher: Copernicus GmbH, 2022.

Ma, K., Feng, D., Lawson, K., Tsai, W., Liang, C., Huang, X., Sharma, A., and Shen, C.: Transferring Hydrologic Data Across Continents – Leveraging Data-Rich Regions to Improve Hydrologic Prediction in Data-Sparse Regions, Water Resources Research, 57, https://doi.org/10.1029/2020wr028600, publisher: American Geophysical Union (AGU), 2021.

Martel, J.-L., Arsenault, R., Turcotte, R., Castañeda-Gonzalez, M., Brissette, F., Armstrong, W., Mailhot, E., Pelletier-Dumont, J., Lachance-Cloutier, S., Rondeau-Genesse, G., and Caron, L.-P.: Exploring the ability of LSTM-based hydrological models to simulate streamflow time series for flood frequency analysis, https://doi.org/10.5194/egusphere-2024-2134, 2024.

Mbouopda, M. F., Guyet, T., Labroche, N., and Henriot, A.: Experimental study of time series forecasting methods for groundwater level prediction, https://doi.org/10.48550/arXiv.2209.13927, arXiv:2209.13927 [cs], 2022.

McKinney, W.: Data Structures for Statistical Computing in Python, in: Proceedings of the 9th Python in Science Conference, edited by van der Walt, S. and Millman, J., pp. 56–61, SciPy, Austin, Texas, https://doi.org/10.25080/Majora-92bf1922-00a, 2010.

Nayak, P. C., Rao, Y. R. S., and Sudheer, K. P.: Groundwater Level Forecasting in a Shallow Aquifer Using Artificial Neural Network Approach, Water Resources Management, 20, 77–90, https://doi.org/10.1007/s11269-006-4007-z, 2006.

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., and Gupta, H. V.: What Role Does Hydrological Science Play in the Age of Machine Learning?, Water Resources Research, 57, https://doi.org/10.1029/2020wr028091, publisher: American Geophysical Union (AGU), 2021.

Ohmer, M., Liesch, T., Habbel, B., Heudorfer, B., Gomez, M., Clos, P., Nölscher, M., and Broda, S.: GEMS-GER: A Machine Learning Benchmark Dataset of Long-Term Groundwater Levels in Germany with Meteorological Forcings and Site-Specific Environmental Features, Earth System Science Data Discussions, https://doi.org/10.5194/essd-2025-321, in review, 2025.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Édouard Duchesnay: Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 12, 2825–2830, 2011.

Tran, V. N., Nguyen, T. V., Kim, J., and Ivanov, V. Y.: Technical note: Does Multiple Basin Training Strategy Guarantee Superior Machine Learning Performance for Streamflow Predictions in Gaged Basins?, https://doi.org/10.5194/egusphere-2025-769, 2025.

Usman, M., Waqar, M., and Ng, C. W. W.: Groundwater level prediction using MIMO-LSTM, 2023.

van Rossum, G.: Python Programming Language, https://www.python.org/, 1995.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., and van Mulbregt, P.: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, Nature Methods, 17, 261–272, https://doi.org/10.1038/s41592-019-0686-2, 2020.

Wunsch, A., Liesch, T., and Broda, S.: Deep learning shows declining groundwater levels in Germany until 2100 due to climate change, Nature Communications, 13, https://doi.org/10.1038/s41467-022-28770-2, publisher: Springer Science and Business Media LLC, 2022a.

Wunsch, A., Liesch, T., and Broda, S.: Feature-based Groundwater Hydrograph Clustering Using Unsupervised Self-Organizing Map-Ensembles, Water Resources Management, 36, 39–54, https://doi.org/10.1007/s11269-021-03006-y, publisher: Springer Science and Business Media LLC, 2022b.

Yu, Q., Tolson, B. A., Shen, H., Han, M., Mai, J., and Lin, J.: Enhancing long short-term memory (LSTM)-based streamflow prediction with a spatially distributed approach, Hydrology and Earth System Sciences, 28, 2107–2122, https://doi.org/10.5194/hess-28-2107-2024, publisher: Copernicus GmbH, 2024.

Zhang, Z., Wang, D., Mei, Y., Zhu, J., and Xiao, X.: Developing an explainable deep learning module based on the LSTM framework for flood prediction, Frontiers in Water, 7, https://doi.org/10.3389/frwa.2025.1562842, publisher: Frontiers Media SA, 2025.

Zhou, Y., Zhang, Q., Bai, G., Zhao, H., Shuai, G., Cui, Y., and Shao, J.: Groundwater dynamics clustering and prediction based on grey relational analysis and LSTM model: A case study in Beijing Plain, China, Journal of Hydrology: Regional Studies, 56, 102 011, https://doi.org/10.1016/j.ejrh.2024.102011, publisher: Elsevier BV, 2024.