Summary

The paper compares single-well models with global models for groundwater level forecasting, focusing on robustness and predictive performance. The comparison is well motivated by earlier work suggesting that global approaches often perform better in surface water modelling. The authors also examine how global models performance depends on training-set size and evaluate the influence of dynamic similarity across sites. The study also investigates how well global models generalize to unseen wells. Overall, the manuscript is clearly structured and the analysis is presented carefully and transparently.

- We thank the reviewer for the careful reading of our manuscript and for the constructive and detailed comments. We appreciate the positive assessment of the study design and the clarity of presentation. The suggestions helped us to improve the manuscript by clarifying several methodological aspects, strengthening the supporting references, and improving the precision of terminology and definitions throughout the text. Below, we provide a point-by-point response to all comments and describe the corresponding changes made in the revised manuscript.

Evaluation and Recommendations

Model choice may influence the conclusions, but it is currently unclear to what extent. For single-well models, performance can vary across sites depending on the selected model structure. Global models performance may also be sensitive to model choice, which could affect the resulting predictions and the strength of the conclusions. Expanding the set of tested models may be beyond the scope of this paper, but I recommend explicitly discussing how sensitive the main findings are to the chosen model(s), and under which conditions the conclusions might change.

- **Response:** We agree that the conclusions depend on the chosen model class, because architecture choice can change absolute skill and thus the magnitude of local–global differences. A comprehensive multi-architecture benchmark is beyond the scope of this study. To keep the comparison controlled, we therefore use fixed benchmark architectures (Ohmer, 2025) identified through preliminary baseline benchmarking (CNN for single-well models; LSTM for global models).
  We added an explicit discussion clarifying that our main results are primarily intended to compare training strategies and training-data composition. We also outline conditions under which the conclusions could change, most notably if architectures with stronger entity awareness (e.g., attention-based or graph-enhanced models) reduce site-specific mismatch in heterogeneous settings and thereby narrow the local–global gap.

  We (i) justify the fixed-architecture design in the Methods and (ii) add a Discussion paragraph that delineates the scope and specifies when alternative model classes could alter the results.

  As an additional diagnostic, a map showing the spatial distribution of performance differences (e.g., ΔNSE = NSE_global − NSE_local) would be informative to assess whether the largest deltas follow any geographic or hydrogeological patterns.
- **Response:** Thank you for this suggestion. We added a new spatial diagnostic figure showing per-well performance differences as ΔNSE = NSE_G − NSE_S. The new figure 1 includes (a) a pointwise ΔNSE map and (b) a LISA cluster map (kNN, k = 8, $p < 0.05$) to identify statistically significant local clusters and spatial contrasts, with hydrogeological regions shown in the background for context.

We reference and briefly discuss this figure in the Results (RQ i), where we show that ΔNSE is close to zero on average but exhibits significant spatial autocorrelation and localized clustering (see Fig. 1 in the manuscript).

The methodology of filtering out a subset of wells is clear and coherent, and the correlation-based selection is easy to follow. However, I wonder how the results might change if a spatio-dynamic clustering were used instead. In this context, it would help to justify why a correlation-based approach was preferred over other clustering methods. A useful discussion point is whether adding hydrogeological classifications (in addition to the dynamic similarity) could provide meaningful context before applying the global model, and whether longer time series (where available) would be expected to improve model performance.

- **Response:** We agree that spatio-dynamic clustering could be a plausible alternative and may lead to different subsets. We chose a correlation-based selection because it is simple, transparent, and parameter-light, and because it provides a continuous and directly interpretable ranking according to our definition of dynamic similarity (i.e., similar temporal groundwater-level dynamics). This also enables a controlled, stepwise reduction (P0 → P5) while remaining directly comparable to the random-reduction baseline.

  We deliberately did not impose a spatial constraint on the similarity criterion, as dynamically similar groundwater responses are not necessarily local, and preserving such non-local analogs is part of the motivation for global learning. While spatio-dynamic clustering represents a plausible alternative, it introduces additional design choices (e.g., spatial weighting and cluster definition) and would make the controlled stage-wise comparability of the progressive reduction (P0–P5) less straightforward. We added a brief clarification of this rationale in the Methods section (Section 3.1).

- Regarding hydrogeological classifications, we already include a broad set of static site attributes that capture hydrogeological context. However, as discussed in the manuscript, these descriptors may still be insufficient to fully explain heterogeneous, site-specific groundwater responses. Concerning record length, longer time series can in principle be beneficial, but we used a standardized multi-decade period across wells to ensure a consistent and comparable experimental design. A dedicated sensitivity analysis of record length is therefore left for future work.

Line 16: missing reference.

- **Response:** Thank you for pointing this out. We agree that this statement should be supported directly at that location, and we added the missing citation by including representative examples (e.g., Kratzert et al., 2019a, 2024; Nearing et al., 2024) to substantiate the claim.

Line 17: are often slower (not always, as in the case of Karst)

Response: Thank you for this important clarification. We agree that groundwater responses are not always slower (e.g., karst systems can react rapidly). We revised the wording accordingly by changing “are slower” to “are often slower” and added a brief clarification in the text.

104-107: Please rephrase for clarity. In the context of this sentence, it is not clear what “unseen location” means

- **Response:** Thank you for pointing this out. By “unseen location” we refer to monitoring wells (sites) that were entirely withheld from model training and used only for evaluation, i.e., spatial out-of-sample prediction at unseen sites. We rephrased the sentence accordingly to make this explicit.

Line 122: was HYRAS or ERA5-Land used in this case?

- **Response:** Thank you for the clarification request. In this study, we used the meteorological forcing exactly as defined in the GEMS-GER dataset: HYRAS-based variables were selected whenever available (higher spatial resolution), and ERA5-Land was only used to complement variables not provided by HYRAS. We revised the text around Line 122 to state this explicitly.

Line 193- 195: “Groundwater drought” is defined and interpreted in different ways across the literature. In this manuscript, it appears to be implicitly defined as periods when groundwater levels fall below the 10th percentile (“the 10th and 90th percentiles of the observed distribution in the test set.”), but this threshold is not stated clearly or justified. Please explicitly define the drought criterion, provide a reference (or brief background) for the use of the 10th-percentile threshold, and clarify your terminology.

- **Response:** Thank you for this important clarification. We agree that the term “groundwater drought” is defined and interpreted in different ways across the literature. In our study, we do not aim to identify drought events in a hydrological sense, but to evaluate model behavior under rare low/high groundwater-level conditions that are extrapolative with respect to the training range.
  We revised the manuscript to explicitly define the criterion as follows: for each well, low extremes are weeks in the test period where observed groundwater levels fall below the 1st percentile of the well-specific training distribution, and high extremes are weeks where they exceed the 99th percentile. We report errors restricted to these subsets. We justify the 1st/99th percentile thresholds as non-parametric, site-specific cutoffs that isolate the most extreme tails while remaining comparable across heterogeneous wells and avoiding assumptions about absolute drought levels.
  Finally, to avoid ambiguity with broader drought definitions, we clarified the terminology by referring to these conditions as “low/high extremes” (or “extrapolative conditions”) rather than implying a general groundwater-drought definition. We also corrected the earlier wording that mistakenly referenced the 10th/90th percentiles to ensure full consistency between the Methods and the implemented evaluation.

How these lines relate to line 305: “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”.?

- **Response:** Thank you for noting this inconsistency. Line 305 reflects the intended and implemented definition of extremes (below the 1st percentile / above the 99th percentile of the well-specific training distribution). The earlier passage referred mistakenly to the 10th/90th percentiles of the test distribution. We corrected this text so that both sections now consistently use the 1st/99th percentile (training-based) criterion.

Section 4.4 is duplicated to 4.5.

- **Response:** Thank you for noticing this. Section 4.4 was duplicated as Section 4.5 in the submitted version. We removed the duplicate section in the revised manuscript.

“Never Train a Deep Learning Model on a Single Well? Revisiting Training Strategies for Groundwater Level Prediction” by Ohmer and Liesch presents an interesting study on the design of DL models for groundwater timeseries modelling. Even though there already exists a substantial amount of DL applications on groundwater timeseries modelling, I believe that the study design and the obtained results add novelty to the existing work. I have several points that I wish to see addressed prior to publication.

- Thank you for your careful and constructive review. Your comments helped us clarify key methodological details (stage definitions, evaluation sets, and normalization), improve the presentation of our transferability experiments, and sharpen the interpretation and positioning of our findings within the groundwater and hydrology literature. As a result, the revised manuscript is substantially clearer and stronger.

In the introduction, the authors give a quite broad overview of DL application for both surface water and groundwater timeseries modelling. The introduction would benefit from clearly stating which studies are focusing on groundwater and which on surface water. Since this is a groundwater study I wonder how many surface water references are required – maybe some of them can be remove and replaced by groundwater references. I agree that there are more DL experiences in the surface water domain, especially when it comes to spatial transferability, but this can maybe also be an additional point to be highlighted in the introduction. The intercomparison study by Collenteur et al (https://doi.org/10.5194/hess-28-5193-2024) would be a good addition to the introduction.

- **Response:** Thank you for pointing this out. We agree that the references previously cited around Objective 4 were predominantly from the surface-water domain. We used these studies primarily to motivate the concept of spatial transferability, as systematic large-sample evidence on cross-site transfer is currently more established for streamflow.
  In the revised Introduction, we now explicitly distinguish between surface-water and groundwater studies and streamlined the surface-water references accordingly. We added groundwater-specific references addressing spatial generalization in groundwater head prediction, including prediction at ungauged locations/areas and out-of-sample evaluations (e.g., Heudorfer et al., 2024; Haaf et al., 2023; Patra et al., 2023), and included the intercomparison study by Collenteur et al. (2024).
  We also revised the wording to avoid an unqualified “first” claim and instead emphasize that systematic large-sample assessments of spatial transferability in groundwater remain scarce. Objective 4 therefore provides a dedicated, large-scale evaluation to help address this gap.

The authors carry out a spatial transferability study (4. objective), which I have not seen in the groundwater literature and the presented references (l.50) are all surface water studies. If this is the first spatial transferability study in the groundwater domain, the authors should state this clearly and if other studies exist, they should be mentioned in the introduction.

- **Response:** Thank you for this comment. We agree that the references cited around Objective 4 were previously focused on surface-water studies. We revised the Introduction to clarify that these studies are used to motivate the concept of spatial transferability (PUB/ungauged basins), while systematic large-sample evidence for spatial out-of-sample generalization in groundwater head prediction remains limited.
  We added groundwater-focused references that explicitly address spatial generalization across withheld wells (e.g., Heudorfer et al., 2024), and we clarified that Objective 4 evaluates leave-well-out spatial out-of-sample prediction.

What is the reasoning behind using a CNN for the single well models and a LSTM for the global models?

- **Response**: Thank you for this question. Our primary objective is to compare training strategies (single-well vs. global vs. partitioned) rather than to benchmark model architectures. We therefore selected model backbones pragmatically, guided by initial screening experiments and computational feasibility.
  For the single-well setting, we tested both CNN- and LSTM-based variants and found that a 1D CNN achieved comparable (and in some cases slightly better) predictive skill while being substantially faster and more stable during optimization. In contrast, single-well LSTM training was more prone to unstable runs (e.g., divergence or exploding gradients), which is critical when training thousands of independent site-specific models. We therefore adopted the CNN as a robust and computationally efficient backbone for the single-well experiments. This observation is consistent with previous findings (Wunsch et al., 2021).
  For the global and partitioned settings, we used an LSTM because gated recurrent models are a widely adopted baseline for multi-site hydrological sequence learning and are designed to capture long-term dependencies. This aligns with the motivation of global training, which relies on information sharing across sites, and is consistent with previous large-sample hydrological studies (e.g., Heudorfer et al., 2024).
- More generally, recurrent models tend to benefit from larger amounts of data per entity, whereas the single-well setting can be comparatively data-limited. Under such conditions, simpler feed-forward sequence models can be competitive or preferable. Accordingly, our conclusions focus on how predictive performance changes with the training-data configuration while keeping inputs, data splits, and evaluation procedures consistent across experiments.
- We added a brief clarification of this rationale in the Methods section (Model Architectures) to emphasize that the study targets training strategy effects rather than an architecture benchmark.

Section 2.2 Are any of the climate variables aggregated in time, for example running sum of net precipitation or SPI at different aggregation windows?

- **Response:** Yes. All dynamic variables are used at weekly resolution exactly as provided in GEMS-GER; daily data were aggregated using the variable-specific weekly operator (mean or sum; see Table 1 in Ohmer et al., 2025). We did not derive additional running sums or multi-window indices such as SPI. We stated this explicitly in Section 2.2 and refer to Table 1 in Ohmer et al. (2025) for the variable-specific aggregation operator.

Section 2.3 Just to be clear, timeseries statistics such as mean head or standard deviation are not part of the static attributes?

- **Response:** Correct. The static attributes are purely site descriptors and do not include any statistics computed from the groundwater-level time series (e.g., mean head or standard deviation); we clarified this in Section 2.3.

What are the sensitivities of the choice of architecture and hyperparameter values presented in section 3.2?

- **Response:** Thank you for this question. We did not perform an extensive hyperparameter optimization or a dedicated architecture sensitivity study because the main objective of this work is to isolate the effects of training strategy and training-data composition. To ensure a controlled and comparable setup, we adopted the benchmark architectures and

hyperparameter settings from the GEMS-GER workflow (Ohmer et al., 2025) and kept them fixed across all partition stages; we clarify this explicitly in Section 3.2.

This choice is consistent with common practice in large-sample hydrological DL studies, where hyperparameters are often taken from established baselines or prior work rather than optimized per site or per experiment (e.g., default/baseline settings, pragmatic choices, or reuse of settings from earlier studies). We do, however, address robustness to stochasticity by evaluating an ensemble of ten independently initialized models and reporting the median performance. In addition, preliminary screening experiments motivated the CNN (single-well) versus LSTM (global) backbone choice, balancing predictive skill with computational feasibility and training stability.

Section 3.2 Were the head timeseries normalized in any way? If yes, how can the authors argue for testing spatial extrapolation if knowledge on mean and standard deviation is required for the back transformation?

- **Response:** Thank you for the clarification request. Yes, groundwater head time series were normalized using a site-specific z-score transformation, computed only from the pre-test period (train + validation) to ensure numerically stable optimization. Predictions were subsequently back-transformed using the same site-specific mean and standard deviation.

  Regarding spatial out-of-sample (OOS) evaluation: in our setup, "spatial OOS" means that the target wells were completely withheld from training the model weights (i.e., no target-well samples were used during training). The normalization/back-transformation parameters for the target wells are derived exclusively from their historical (pre-test) observations and therefore do not use any information from the test period.

  Scope/limitation: This setting evaluates the transfer of learned model weights to previously unseen wells with historical head records (needed for well-specific scaling), not prediction at sites without any head observations. A fully "no-head-data" setup would require avoiding site-specific target scaling (e.g., global training-based scaling, a baseline–anomaly formulation, or a short initial calibration window to estimate offset/scale). We clarified this definition and limitation in the manuscript (Section 3.2 / Experimental Design).

What does P1, P2, …, P5 mean? P1 excludes 500 wells, P2 excluded 1000 wells, and so on? To me this first became clear when reading the result sections. It would be good to state the number of wells in each stage already in 3.1. The testing strategy is not stated. Are all wells for 2013-2022 used for testing or only the ones left after stagewise removing?

- **Response:** We agree that the meaning of stages P1–P5 should be stated more explicitly earlier in the manuscript. We therefore clarified in Section 3.1 (Methods) that partitioning stages are cumulative: starting from P0 (n = 2951), each stage removes an additional 500 wells from the previous stage (P1: 2451; P2: 1951; P3: 1451; P4: 951; P5: 451).

  We also specify that COR removes the lowest-representativeness wells based on mean absolute correlation, whereas RND removes wells at random with the same size progression. For each partitioning stage (and for COR/RND separately), the model is trained, validated, and evaluated only on the wells remaining in that stage. The temporal test window (2013–2022) is kept fixed, but the evaluated well set changes with the stage (P0–P5).

From Figure 2 I get the impression that the testing dataset varies for stage – can the performances be compared in a meaningful way across the stages? I would suggest to make an additional test using the P5 wells for all stages.

- **Response:** Thank you for this important point. Following your suggestion, we added an additional like-for-like evaluation in which all stage models are assessed on the same fixed well set, namely the final P5 wells. This ensures that cross-stage differences are not conflated with changes in the evaluated well population.

  The revised analysis is summarized in Fig. 6, which distinguishes two complementary perspectives:

  Fig. 6a (stage-dependent): ΔNSE between successive retrained global models (e.g., P2 → P3), computed on wells present in both consecutive stages (intersection set).

  Fig. 6b (fixed test set): ΔNSE(Px → P5) = NSE(P5) − NSE(Px), evaluated on the same P5 wells for all comparisons (constant test set), as requested.

  During this revision, we also aligned the ΔNSE definition with the intended interpretation: the previous visualization reported differences relative to P0, whereas the discussion referred to stage-to-stage changes (as shown in Fig. 6a). We therefore replaced the original plot (Fig. 5) with the updated Fig. 6a/b formulation, which removes ambiguity and directly addresses the reviewer's concern.

That fact that global models do not outperform single well models for the P0 stage and that an advantage of global model first becomes tangible at P4 and P5 makes me wonder if the chosen LSTM architecture can exploit the static features in a meaningful way? Along these lines, when removing wells based on their correlation, do the static features also become more homogeneous? In other words, is the similarity of the timeseries reflected by the static features?

- **Response:** Thank you for this thoughtful question. We agree that the fact that global models do not clearly outperform single-well models at P0, while advantages emerge mainly at later COR stages (P4–P5), is consistent with limited "entity awareness" under strongly heterogeneous groundwater dynamics. Although the global LSTM uses dynamic inputs plus >50 time-invariant static attributes (hydrogeological, topographic, soil, land-use), these descriptors do not include statistics derived from the groundwater-level time series and may still be insufficient to uniquely identify the controlling local conditions that drive site-specific responses (e.g., fine-scale geology, flow paths, abstraction).

  As a result, wells that appear similar in static feature space can exhibit distinct dynamics, and a global model may tend to learn more generic, averaged behaviors rather than reliably selecting the correct regime for each site. Our partitioning results support this interpretation: correlation-based removal reduces the presence of dynamically atypical wells and thereby decreases conflicting learning signals, making cross-site learning increasingly effective and performance changes more consistent across wells. Random removal largely preserves heterogeneity and therefore yields no systematic gains.

  While it is plausible that static features become somewhat more homogeneous under COR, the removal criterion is purely dynamic (and not spatially constrained), so the observed

improvements are best explained by increased dynamic consistency rather than by a trivial homogenization of static attributes. We cannot exclude that alternative architectures could exploit static features more strongly, but given the already rich descriptor set, we interpret the main limitation as the incomplete observability of key local controls in large-scale static datasets rather than model capacity alone.

In the revised manuscript, we explicitly reflect this interpretation in the Discussion (Comparison with previous studies) and the Conclusion by emphasizing limited entity awareness under heterogeneous dynamics and by clarifying that correlation-based filtering acts on time-series similarity rather than on static-feature similarity.

Another very relevant question in my opinion is the length of the timeseries. The authors make use of an extensive German database, with full coverage for a period of 1991 to 2022, which is a coverage that is not available in many other countries. Therefore, an alternative modelling experiment with stages where e.g., 2 years at a time are removed from the training dataset would be very insightful. P1 starting in 1996, P2 in 1998, etc. would be extremely relevant for similar applications in countries with shorter groundwater records.

- **Response:** Thank you for this relevant suggestion. We agree that sensitivity to record length is important for assessing transferability to regions with shorter groundwater records. Following this suggestion, we implemented a progressive start-year experiment in which the training period is shortened stepwise by removing the earliest years while keeping the evaluation period fixed.
  The results of this analysis are summarized in the new Fig. 2. Panel (a) shows the median NSE as a function of training-set length for single-well (S) and global (G) models, while panel (b) illustrates the corresponding distributions of NSE values across wells. This experiment allows us to systematically evaluate how predictive performance changes as the available training record decreases.
  We discuss these results in the Results and Discussion sections and relate them to previous studies investigating the influence of record length on model performance. The implications for transferability to regions with shorter monitoring records are also briefly reflected in the Conclusions..

Section 4.4 and 4.5 contain the same text.

- **Response**: Thank you for pointing this out. We removed the duplicated paragraph, so Sections 4.4 and 4.5 no longer contain the same text.

I am puzzled why the performance for the correlation wise stages increases in figure 6 for the out of sample wells. For P5cor the model is trained on 451 and tested on 2500, and for P1cor the mode is trained on 2451 and tested on 500, is this correct? Again, the varying testing datasets make it difficult to compare performance across the stages in my opinion. Nevertheless, for the P5cor training you are using very homogeneous timeseries and test it across very heterogenous timeseries. Why should this work better than P1cor where you are training using heterogenous timeseries and also use heterogenous timeseries for testing?

- **Response:** Thank you for this important observation. In Fig. 8 (formerly Fig. 6) we evaluate out-of-sample (OOS) performance on the wells excluded under each stage. Under correlation-based exclusion, the removed wells differ systematically across stages: early stages (e.g., P1_COR) remove the most dynamically atypical wells (lowest mean absolute correlation), whereas later stages remove wells that are progressively more similar to the

remaining training pool. Consequently, the OOS test set is stage-dependent by design and becomes “less dissimilar” on average with increasing stage.

Regarding sample sizes: the training set size decreases cumulatively (e.g., P1_COR: 2451 wells; P5_COR: 451 wells). The OOS evaluation at stage PX is performed on the wells removed up to that stage (cumulative excluded set; e.g., n = 500 at P1 and n = 2500 at P5), i.e., the evaluated well population changes across stages. Therefore, strict cross-stage comparisons should be interpreted as transfer to different target populations rather than as a like-for-like comparison.
T
he apparent increase in OOS performance under COR is therefore expected: later-stage OOS targets are, on average, closer in dynamics to the remaining training subset, so transfer becomes easier even though the training set is smaller. This is consistent with the fact that many wells in our dataset share broadly similar dynamics, so that even a small but homogeneous training subset can still represent the dominant behavior patterns. At early stages, targets are highly atypical, and limited entity awareness (static descriptors not fully capturing fine-scale local controls) makes transfer difficult. Under random exclusion, the OOS target distribution is not ordered by dynamic dissimilarity, which explains the more stable behavior across stages.

We clarified this evaluation setup in the manuscript by explicitly describing the stage-dependent (cumulative) OOS evaluation in the Methods and by explaining it in the caption and text of Fig. 8. This ensures that cross-stage differences are interpreted as transfer to different target populations rather than as a fixed-test comparison.