General assessment

The research article “Exploring the generalisation ability and interpretability of Long Short-Term Memory (LSTM) networks for large-sample groundwater level predictions” by Fang et al. (EGUSPHERE-2026-3267), investigates the generalisation and interpretability of multi-station LSTM models across England. The large observational dataset, evaluation at unseen stations, and comparison with a mechanistic model provide a useful contribution to regional groundwater modelling. I recommend major revision, primarily to strengthen the interpretation of the station comparisons and the robustness of the CART and PFI analyses. The comments below seek supporting analyses and clarification of these issues, together with corrections to several reporting inconsistencies.

Major comments

Comment 1: Lines 153-155 and Fig. S3. Figure S3 usefully presents the spatial distributions and dominant fluctuation-cycle proportions of the training and unseen stations. Please extend this comparison to aquifer composition and key environmental attributes, and discuss whether differences in station composition may contribute to the performance gap between the two sets. This would help distinguish the influence of site characteristics from that of predicting at stations excluded from training.

Comment 2: Lines 296-311. Please further explain the rationale for excluding these 15 poorly predicted stations from the CART analysis. Low correlations with neighbouring stations may reflect genuine local variability or limitations in the available environmental descriptors, which may themselves help explain model performance. A comparison of the CART results with and without these stations would help assess the robustness of the main conclusions. Please also clarify the pool of eligible neighbouring stations, the period and minimum overlapping record length used to calculate inter-station correlations, and the treatment of stations without eligible neighbours within 20 km.

Comment 3: Lines 317-319 and 331–334. The manuscript states that permutation across samples and time steps disrupts the spatial and temporal structure of dynamic inputs. Please clarify whether static attributes remain constant within each station and whether values from the 30-day and 270-day aggregation windows are mixed. As these operations may also generate input combinations unlike those used for training, a sensitivity check using a permutation scheme that preserves these relevant structures would strengthen interpretation of the PFI rankings. Please also specify whether grouped PFI uses a shared permutation order for all variables in each group.

Specific comments

Comment 4: Lines 235-237. The cited approach of Kratzert et al. (2019a, Appendix B) partitions basins and performs hyperparameter selection within the calibration period. Please confirm whether the corresponding procedure here splits the 636 training stations into four folds and uses only 1971–2014 data. A brief account of the partitioning and selection criterion would make the implementation clearer.

Comment 5: Abstract, lines 15-20. Please state which LSTM configuration performs better at unseen stations and clarify which models are meant by “both models” in the subsequent benchmark comparison. The statement that NSE alone can be misleading could also distinguish skill in reproducing temporal fluctuations from accuracy in predicting absolute groundwater depth, since a high correlation does not remove the mean and variability errors reflected in NSE.

Comment 6: Lines 59-61. Please reconsider the scope of the statement that only three studies have investigated multi-station LSTM groundwater modelling. Other examples of jointly trained multi-well models include Patra et al. (2023; doi:10.1016/j.ejrh.2023.101442) and Di Taranto et al. (2025; doi:10.3390/w17152268). Their inputs and forecasting settings differ from those of the present study, so either discussing these studies or explicitly defining the narrower class of models intended here would make the literature coverage clearer.

Comment 7: The introduction provides a helpful context for the study. A brief synthesis at the end of how the present work builds on previous studies could further help readers appreciate its main contribution.

Comment 8: Lines 229-233 and 261-263. Please clarify how the random static identifiers were generated, including whether each station was independently assigned a 17-dimensional random integer vector. As ten training seeds are already described, please also specify whether the station identifiers were regenerated between runs or whether only the network initialisation was varied.

Comment 9: Lines 249-250 and Eq. (1). The NSE* notation follows the cited loss-function convention, but describing it as the average station NSE may cause confusion with the evaluation metric in Table A3. Please explicitly distinguish the normalised squared-error training loss, which is minimised, from the evaluation NSE, for which larger values indicate better performance.

Comment 10: Lines 274-283. The comparison over common stations and a shared testing period provides a useful benchmark. Since the LSTM was trained on groundwater level observations whereas DECIPHeR-GW was calibrated against streamflow, a brief discussion of this difference in calibration information would help contextualise their relative performance.

Comment 11: Lines 264 and 329. Please correct two cross-references: the ensemble comparison at line 264 is shown in Fig. S5, not Fig. S4, and the six evaluation metrics at line 329 are described in Sect. 3.2, not Sect. 3.3.

Comment 12: Line 385 and Fig. 3a. The text reports 52 stations, whereas the corresponding branch in Fig. 3a contains 53 (25 + 28). Please reconcile the count and the associated percentage.

Comment 13: Lines 374-383 and 553-554. The CART subset is clearly defined, and the reported 93% is conditional on this selection. Please consider whether the broader conclusions about better performance in high-specific-yield and high-transmissivity aquifers also hold across all 341 out-of-sample stations. Showing that relationship for the full test set, or qualifying the conclusions as findings for the selected subset, would strengthen the interpretation. The phrase “hydrogeological controls” could also more clearly distinguish statistical associations from mechanistic explanations.

Comment 14: Lines 404-408 and Fig. 4. The category “both models perform well” is defined by $r > 0.75$ for at least one model and a between-model difference of less than 0.25. These criteria do not ensure that both models exceed $r = 0.75$. Please consider revising the category label or requiring both models to meet the stated performance criterion, and clarify whether Δr denotes the absolute difference.

Comment 15: Lines 426-446 and Figs. 5, S13 and S14. The supplementary figures already provide the importance ratios. Please draw on these values when interpreting the rankings in Fig. 5 and summarise the variability across the ten permutation repeats. A brief explanation of ratios below 1 would also help distinguish negligible or potentially adverse feature contributions from consistently beneficial ones.

Comment 16: Lines 438-440: Figure S12 presents correlations between long-term meteorological means across stations, whereas the explanation here concerns shared seasonal information within time series. Please distinguish spatial from temporal correlations and either provide supporting temporal analysis or present this explanation as a possible interpretation.

Comment 17: Lines 355 and 478. The proportion of LSTM_RND stations achieving Pearson $r > 0.75$ is reported as about 20% in the results and about 15% in the discussion. Please check and reconcile these percentages.