General comment: The study presents the first (or one of the first) global/regional deep-learning based groundwater model for the UK. The results are very well worked out throughout, language and presentation is excellent, and the content represents valuable contribution to the scientific body. I do not have major comments and propose publication after minor comments (below) are addressed.

Line 59-69: There is now also a global LSTM in Australia https://doi.org/10.5194/egusphere-2026-2950

Line 71-72 you write: "Again, they found that LSTM models using random integer features performed comparably to those employing meaningful environmental attributes." Please be aware of the corrigendum (https://doi.org/10.5194/hess-28-525-2024-corrigendum) to our 2024 article you cite here. Due to a coding error that we sadly discovered only a year after publication, the OOS model runs were slightly off. In the corrected version of the model, the meaningful environmental features do outperform other variants – however only by a little. Your wording "comparably" is vague enough to also capture the corrigendum results as well, so I don't know if you are already aware of the corrigendum or not.

Line 73-75 you write: "These findings imply that the LSTM model utilises static features primarily as 'unique identifiers' to memorise local station behaviours during training rather than deriving the generalisable hydrological insights required for spatial extrapolation." It is important to note that this is not proven, but a hypothesis based on these results. We expanded the discussion about this hypothesis in the 2025 paper (https://doi.org/10.1029/2024GL113036) and came to the conclusion that the model with meaningful static features does learn hydrological viable information, especially considering the forest cover reduction experiment in the appendix. This experiment was in response to backlash from some scientists who strongly argued that performance-based indication are no proof with regard to what the model does "under the hood", which is why we ran the forest cover reduction experiment. Nowwe did not compare this with random feature-driven model in the 2025 paper, and of course these results were derived in hydrology and not hydrogeology, where static features are weaker predictors for hydrogeologic system functioning. But from the collective results currently I wound say that the "unique identifier" hypothesis might hold for random-feature based models, but not or only partially for models based on meaningful static features.

Line 78-80: That is the correct conclusion of this section in my opinion.

Line 94-08: 100% agree.

Line 115-116: May be true in published form; Andreas Wunsch and Tanja Liesch made an extensive comparison in the upper Rhine valley ca. 2020/2021, but only available as project reports.

Line 147-149: All of the data selection criteria in 2.1 are reasonable and quite common. But I want to highlight that data availability sub-selections are **unnecessary from a strictly technical standpoint** in regional DL modelling. From a technical standpoint, even single groundwater measurements in a given location can be thrown in a sample bag as long as input feature space is gapless (as e.g. done in https://doi.org/10.5194/hess-28-1215-2024). But actually even gaps in input space are not necessary exclusion criteria, but can be addressed with appropriate architectural modifications (https://doi.org/10.5194/hess-29-6221-2025). For target feature availability, comparatively short time series are still sensible to include in your dataset

(https://doi:10.2166/nh.2026.119 for discussion of this). I am aware that the present study is already run, and there is nothing strictly wrong about this sub-selection. But keep that in mind for next time, before unnecessarily limiting your data set. Because it looks like you really drastically reduce your dataset in this step (1384-->636).

Line 177-179: In this context, I can highly recommend reading https://doi.org/10.1111/j.1745-6584.2005.00090.x

Line 217: Tmin+Tmean+Tmax has generally more information that Tmean+Trange.

Line 219-227: Great approach, and there is formal empirical evidence from our lab that this works excellently (https://doi.org/10.5194/hess-29-1749-2025). My only comment is that considering your target resolution is daily, the last couple of days should be input in daily resolution. But again, not strictly wrong, I don't see a necessity to change it for the current study.

Line 254-255: That is incorrect. The reason station-averaged NSE is used in training is to account for differences in target value location and magnitude (flow height in streamflow, water table amplitude in groundwater). The original reason for introduction is that larger catchments have larger streamflow values, and this value magnitude disbalance remains after standardizing globally before training (as you also do), leading to disproportionate influence of large catchments on the loss, and therefore via the gradients on weight update. Hence learning is biased towards larger catchments. The catchment-averaged NSE mediates that, allowing equal degree of learning from differently sized catchments. In groundwater, not only the magnitude of the time series, but also the overall mean experiences large shifts. This is why we advocate for local scaling instead of global scaling ahead of training in our 2024 paper (see end of chapter 3.1 in https://doi.org/10.5194/hess-28-525-2024) combined with MSE as loss, but that comes at the trade-off of lack of ability to do proper OOS, which is why the scaling approach you use is correct, strictly speaking.

Figure 2/Chapter 4.1: you should report at least the median for each of the models (lines) depicted in figure 2/ discussed in chapter 4.1. If possible, including some uncertainty measure.

Chapters 4.2-4.4: excellent analysis, nothing significant to add/comment.

Line 457-458: That is too strongly formulated and therefore incorrect as such; see my comment to lines 73-75 above.

Line 468ff: Ah yes, there it is properly addressed, great.

Line 482ff: very insightful discussion of the NSE limitations.

Line 527-530: I agree that this would be very very valuable.

Table 1: I have a small problem with the informality of the table, since there are no proofs provided within the table; I agree to all of these points and agree that it adds value to the paper; also I realize some of these points are made implicitly by the results of your paper, while other points are commonplace and generally accepted (e.g. using multiple metrics). It would be good if another column is added called "justification" or "proof" (the latter maybe too strong), where you point to citations or results of your own paper underscoring the points made. Also it would be better to rename column "checklist" to "lesson".