Dear Editor,

Thank you for your constructive feedback. We appreciate your insightful comments and feedback for improving our manuscript. Please find our point-by-point responses below in bold; the original comments are in italics.

The authors have considered the main critical issues raised by the reviewers during the revision process. For this reason, while in my previous assessment I was inclined in requiring a further revision process, after a thorough analysis of the manuscript I decided not to go for a second round of review.

This been said, upon a careful review of the manuscript, I must highlight that it still contains a significant number of typos and awkwardly constructed sentences that require the attention of the authors. While I understand the time constraints, especially with one of the authors being a Ph.D. student nearing graduation, it is crucial to uphold the quality of the manuscript for the benefit of both the journal and yourselves for the purpose of effectively communicating your research. Below, I have highlighted few (non-exhaustive) examples of suggested changes to be made.

Thank you once again for your valuable feedback and for allowing us to enhance the quality of our manuscript. We have now addressed all the suggested corrections as below.

Line 25 " …to learn the dominant station …. preferentially…", I suggest to rephrase the sentence as "…to preferentially learn…"
Updated

Line 37 "…for grasping a more global view of water reserves.." Here the adjective "more" is not appropriate. I suggest rewriting as "for grasping a global view of water reserves"
Updated

Line 43 "Building the large -scale model…..." should be "Building a large -scale model…"
Updated

Line 46 "However, the numerical, physics-based representation of all the physical processes occurring during the hydrological cycle ….." The sentence is awkward; consider rewriting it, for example as "However, the numerical, physics-based, representation of all processes occurring during the hydrological cycle ….
Updated

Lines 160-161 and lines 168. The repetition of the sentence "All the wells considered in the study are in unconfined aquifers" should be rectified.

Corrected

Line 188 "Gualtieri (2022) highlighted that ERA5 uncertainties were greater in mountainous and particularly in coastal locations ….". Consider rewriting this sentence, for example as: "Gualtieri (2022) highlighted that ERA5 uncertainties are greatest at mountain and coastal locations …..

Updated

Line 193 "However, for our study area, we have been evaluating different potential alternative reanalysis products, such as….". Consider rewriting this sentence, for example as "For our study area, we evaluated several potential alternative reanalysis products, such as ….."

Updated

Line 205 "BDLISA was originally designed at a 25km scale and later upscaled to larger scales. For our study, we kept information coming from BDLISA at its original scale (25km), which means aquifer static attributes have a resolution of 25km".

Since you are using the original scale (without upscaling), I would remove the sentence "For our study…", to avoid confusing the reader.

Removed

Line 235 "… we have to make sure" replace by "…we need to ensure…"

Updated

Page 9. Remove the Figure without number and caption, it is already included in Fig. 1.

This figure was only visible (with striked out) in previous tracked version only to represent the change but this is now avoided in the current revision

Line 238 "Also, for some attributes like hydraulic conductivity, it might not be straightforward to get the most relevant resolution, which is needed to account for the most appropriate characteristic describing the well…". Please consider rephrasing this sentence, did you mean "describing the aquifer"?

Updated

Line 272 "Details of range of hyperparameters used are shown in Table 1". The correct reference should be Table 2

Updated

Line 379 "….Furthermore, to facilitate hyperparameter tuning, the last 20% of the training data was used as a validation set" and line 385 "This split corresponds to approximately 80% of 385 the data for training and 20% for testing." These two sentences convey the same information. Please rewrite.

Updated

Line 445 – 446.Clarify the basis for stating "While single station models perform best" as it is not evident from Fig. 8-9

Figures 8 to 10 show the best GWL simulations obtained of different types (annual, mixed and inertial) for single and multi-station models. For those particular cases, both approaches perform similarly and lead to good performance. However, the single-station seems to perform best for inertial GWL type for training by simple visual assessment, and it is clear from the comparison of KGE values of all stations (Fig.7) that the more specialised single-station models generally gave the best results overall, although not significantly. This is more specifically true for inertial GWL, where regional model performances reach the same level as single-station models. While single-station models perform well, multi-station models are valuable when single-station modelling is impractical due to data limitations or computational requirements. For instance, for inertial types where the length of training data might be an issue (e.g. Chidepudi et al., 2024), the performance of the wavelet multi-station models was completely comparable to single-station models (Fig.7, wavelet models/inertial types), showing that in the case of data limitation, the regional approach seems to compensate the lack of temporal depth of available time series.

Updated

Line 470. "In particular…".Consider rephrasing this sentence.

Updated as follows:

"The distribution of data points on the SHAP diagram indicates either a positive (right side on the x-axis) or negative (left side on the x-axis) impact on the output variable. In contrast, the colour scale indicates the range of feature values in which red represents large values, and blue represents small ones of the corresponding feature. Features (input variables) are organised from the most to the least influencing, from top to bottom, based on each feature's mean absolute SHAP values. For instance, in Figure 11a, total precipitation (tp) is the most

influencing feature on the GWL output, and the large feature values on the right (red) correspond to a positive influence on GWL (high GWL with high total precipitation). On the left-side, negative tp SHAP values indicate lower precipitation values contributing to the low GWLs."

Lines 552-560. The following sentence is unclear "In the framework of our study, we decided to exclude some relevant characteristics such as vadose or saturated zone thickness: even when averaged over quite long periods (several years), these values actually represent GWL (the target variable)….. These authors concluded that the models did not show any entity awareness and eventually utilized static attributes as simple identifiers (almost similar to the OHE approach presented herein), meaning that the models did not make use of the relevant (hydro)geological information." Please rewrite.

Updated as :
In the framework of our study, we decided to exclude characteristics such as vadose or saturated zone thickness. Such variables have been used in previous studies (El Janyani et al., 2012 and 2014; Haaf et al., 2023) and considered static (averaged over long periods of time) to investigate the impact of (hydro)geological and geomorphologic characteristics on GWL behaviours. Yet, in our study, it was not relevant to consider such characteristics as "static" since they are linked to the varying GWL which we aim to simulate. Other types of static characteristics reflecting the hydraulic properties of the aquifers, such as hydraulic conductivity, transmissivity, porosity or storativity, were also discarded. While informative in terms of hydrological knowledge, it is likely that: 1- their availability may not be guaranteed over large areas, hence limiting their usefulness. 2- their representativeness as numeric values might be questionable in contexts where spatial heterogeneity is high: in such cases, more general qualitative descriptors such as "fissured" or "porous" might be preferable, as using precise values of hydraulic conductivity, etc., would likely make the models very sensitive to hydraulic heterogeneity which can not be accounted for so precisely. In addition, in a recent and relevant study on "entity-aware deep learning models with static attributes," Heudorfer et al. (2024) highlighted that the models developed did not actually show any entity awareness and eventually utilised static attributes as simple identifiers (almost similar to the OHE approach presented herein), meaning that the models did not make use of relevant and precise (hydro)geological information.

Lines 542-545 "… for alluvial aquifers with probably quite high hydraulic conductivity overall" Clarify the meaning of "probably" here.

Removed the word "Probably" and indicated the median value from the study we were referring to.

Conclusions, line 595-600 "In this article, we introduced the following question: "What's the best way to leverage regionalised information?". In light of our results, it then seems like this is highly dependent on the amount and types of static attributes. It is generally expected that a much higher number of static attribute types would allow for a much better improvement of the multi-station simulation approach. However, Heudorfer et al. (2024) found no improvements using around 28 static features (including 18 environmental and ten time series-based). Also, as pointed out by these authors, employing static attributes for model training might be more relevant in applications on larger scales and/or larger datasets". Please consider rewriting these sentences. As written, it seems that your results are in contrast with the findings of Heudorfer et al. (2024). On the other hand, my understanding is that static attributes can be used for model training on large scales, while they are not particularly useful on small-moderate scales (as such investigated by Heudorfer et al., 2024). Please also include the magnitude levels for both "large" and "small" (or moderate scales.

Our conclusion does not contradict Heudorfer, and this is now clarified in the main text: we found that the most decisive component of improvement was Wavelet and clustering, and then static provides minor additional improv, which can be valuable but, again, not the most decisive.

Updated

Table 4. Fix the typo ("Lattitude")

Corrected

Additionally, I recommend enhancing the quality of Figures 1-2,8-13, several labels are blurred or not visible.

The figures have now been updated with increased sizes of all fonts and image resolutions.