Response to Reviewer 2 of the Manuscript: Training deep learning models with a multi-station approach and static aquifer attributes for groundwater level simulation: what's the best way to leverage regionalised information?

We appreciate the constructive comments by the reviewer. Original comments by reviewer are in italic format and our response is in bold.

Review comments on the manuscript: Training deep learning models with a multi-station approach and static aquifer attributes for groundwater level simulation: what's the best way to leverage regionalised information? by Chidepudi et al.

The manuscript presents several different deep learning approaches to simulate groundwater levels. Dynamic as well as static variables are used to train deep learning models to represent fluctuations on a high temporal resolution (daily data) in northern France. These different deep learning models were combined with different sets of input data (including preprocessing) and training strategies. Overall, the work is timely and covers the important topic of data-driven approaches to simulate dynamic groundwater levels. However, the manuscript has several shortcomings which are listed below. Major revision is needed.

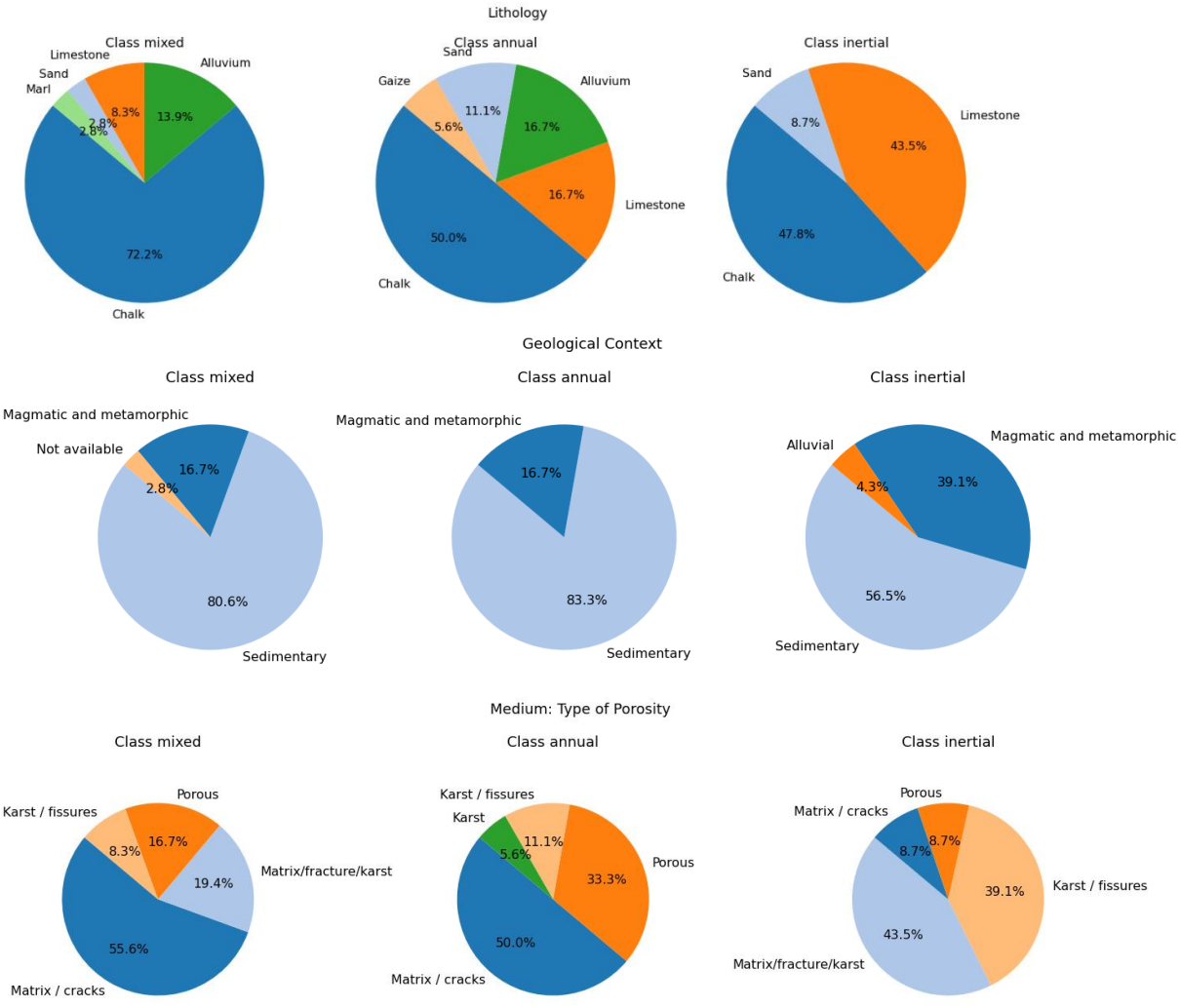
Thank you for taking time to give detailed and constructive comments. We will address all the remaining issues listed in a revised version following our responses below.

Main Comments

What is the best way to leverage regionalised information? - The authors raise this question in the manuscript title but in my opinion, they do not answer the question in a sufficient way. This has mainly two reasons:

- *The manuscript seems to be a combination of a technical note and a case study which leads to the result that a lot of essential information are missing. Reviewer 1 already pointed out several of the technical issues. In addition, a description of the data set is entirely missing. The only information available for the reader is the rough distance between the observation wells and the density in the region. Important information to understand the results and therefore the feasibility of the applied methodology is not supplied by the authors. For example: What is the distribution of static attributes in the different cluster groups? Looking at the attributes presented in Table 1, large differences between lithologies are to be expected (e.g karst vs. clay). Could it be that the annual group consist mainly of observation wells located in karstic/fractured areas and what would this tell us about the outcome of the study? Are these static attributes even presented/discussed in Chapter 4 (I assume that you can see them in Fig. 9 but they are not even named somewhere?*

Technical issues also pointed out by Reviewer #1 mainly concerned the presentation of the hyperparameters eventually selected or optimized, and the architectures of the recurrent-based models. We explained how these comments can be addressed in our response to Reviewer #1. Regarding dataset presentation issue: in the version of the paper submitted, we presented the databases used, including the number, length, duration, and sampling rate of the groundwater level time series, as well as a table of static attributes. Missing information or not provided at the right place, as pointed out by Reviewer #2 (e.g., the number of stations in each class, which was initially presented in the discussion section) will be either completed or moved to the appropriate data section. For instance, we propose to add an in-depth comparison of attributes available for different types of groundwater levels, along with improved details of the datasets, as described below. The three static attributes for different types of groundwater are shown in the pie plots below. However, it is important to keep in mind that such information is always very local and only valid for a given well. A full description of all these attributes will be included in the form of a table in the appendix.



- *The presentation and discussion of the results lacks the already mentioned discussion of the regional context but also a discussion of the results in a broader context. For example, the authors write L398:"However, wavelet pre-processing shifts the importance towards dynamic components, reducing the contributions of static features or OHE. When clustering is combined with wavelet preprocessing, low-frequency precipitation components emerge as key contributors, improving model performance.*

Does this mean that the importance of all dynamic components is higher by default, and we do not need to consider geological/hydrodynamic/topographic features? Does this apply to all kind of unconfined aquifer systems (shallow, deep, karstic…)? Here it would be interesting to combine/compare your results with/to other available publications considering static attributes on a regional scale (e.g. Heudorfer et al., 2024 or Haaf et al., 2023).

In the present paper, we aimed only to assess whether, in our context of relatively parsimonious availability of basin properties, considering such attributes within the framework of DL modeling would significantly improve the simulations. For the sake of the generalization capabilities of DL models, we also probably need to find a reasonable trade-off between the use of all possible/relevant static features and their availability over large areas.

We cannot expect static characteristics to be more important than precipitation, temperature, or other variable time series of the water cycle in explaining groundwater level (GWL) time series variations. As mentioned above, the aim here is to assess to what extent available static attributes, in combination with indispensable forcing hydrological variables, may help refine and improve GWL simulations for stations in various (hydro)geological contexts. This (hydro)geological information is largely accounted for in the weights of the neural network model, but the question remains whether additional static information can be helpful. Our results suggest that in some cases, particularly for the most inertial groundwater level types that mainly record low-frequency, climate-like information, improvements can be gained by adding static features.

We agree that a more thorough comparison with papers that have used static attributes on a regional scale is needed and will add this to the discussion section.

Since the purpose of the paper presented here is not to determine the forcing factors of groundwater level variations, comparison with such state-of-the-art studies will help to put our results into perspective, inasmuch as a comprehensive evaluation of such links would require specific approaches. Such approaches have already been undertaken and presented in numerous previous works that we will use to feed the discussion about this important topic as in (Lee et al., 2019; Heudorfer et al., 2019; Liesch and Wunsch, 2019; Haaf et

al., 2020; Giese et al., 2020). In our own previous works (albeit for the Normandy region only), the linkages between groundwater level variability and potential forcing factors such as the thickness and lithology of surficial formations, aquifer thickness, vadose zone thickness, upstream/downstream location along the flow path, distance to the river, presence of karst, etc., were investigated using dedicated approaches (Slimani et al., 2009; El Janyani et al., 2012, 2014).

The quality in writing (language, clarity etc.) differs a lot throughout the manuscript. This makes it difficult to follow the central theme and therefore requires revision. Sometimes sentences reoccur, e.g. L73: DL models have proved effective on a local scale, and are also on a larger scale by collectively training a significant number of piezometers (Chidepudi et al., 2023b; Heudorfer et al., 2024) vs. L80: The DL models have proved effective at local scale and are also proving more effective on a larger scale. At the same time the introduction of terms and abbreviation is totally off, some examples: GWLs is first introduces in the Introduction and then again in line 185, 308, 378 and 436; SHAP is first introduced in line 231 and then again in 461; an introduction (even though they are quiet common) for AI/DL/KGE and NSE is entirely missing. Altogether it feels like sections/paragraphs of different origin were put together.

We will provide more explanation about KGE and other terms and metrics, and will improve the text with appropriate introduction of terms wherever needed. Also the entire text will be checked for homogenization of the writing quality.

Secondary Comments

*L85: sensitivity to human activities - I do not really understand why this is an **additional challenge compared to runoff data**. Does it mean runoff data are not sensitive to human activities (e.g. river straightening, dam construction etc.)?*

We agree that "additional challenge" was certainly not the most appropriate term. Here we meant to say that groundwater level data are affected by different types of challenges with respect to human activities. This can be confusing and then needs to be modified in the text.

L121: their application to GWL simulation is still questionable. – Do you really mean questionable?

We agree "questionable" is clearly not the right term. Suggested revision: "their application to GWL simulation is still not fully explored or validated across diverse hydrogeological settings."

L141: We refer to (Beven and Young, 2013), for differences in the use of the terms simulation and forecasting. - I do not see the connection between the sentence and the rest of the paragraph. Maybe a few more words are needed?

We updated it as : "We would like to highlight at this point that the present study is not dedicated to ‘forecasting’ as it is the case in most applications of DL to groundwater modeling. The reader can be referred to Beven and Young (2013) for distinctions between 'simulation' and 'forecasting'. In brief, according to their framework, 'simulation' means reproducing system behavior without using observed outputs, while 'forecasting' involves reproducing system behavior ahead of time based on past observations. This study focuses on simulation to understand GWL dynamics, rather than forecasting future levels. This distinction is important for framing our approach and interpreting our results."

L164: Although they seem somehow redundant, they are expected to provide complimentary information about the hydrogeological nature of the hydrosystems – This could and should be tested at one point (which does not mean that you have to add it here).

We agree that it would certainly be interesting to conduct some statistical analysis (multivariate, for instance) to assess the potential redundancy of the information provided by the different static features, but 1- we agree with reviewer #2 that this should probably be undertaken in the framework of one dedicated study (cf. our response to some previous comments), 2) from the DL point of view, redundancy should not be an issue, DL models are basically designed to handle (and learn from) as much information as possible without taking into account any possible redundancy within the data (the model will adjust its parameters according to the most useful information detected). For instance, one part of the useful information can be common to 2 features, and at the same time one other part can be specific to each. It will not be detrimental to the performance of the model. As hydrologists, we only ensured that the input data are hydro-geologically relevant (albeit strictly speaking, from the DL standpoint, the models can even get rid of irrelevant data itself during the learning process).

L167/ L173/180/323: Baulon et al., 2022a/b?

Corrected to Baulon et al., 2022

L187: Bidirectional LSTM - I would be good to provide a reference especially since you write in L192: BiLSTM […] are particularly good at identifying various patterns in data sequences, making them ideal for simulating GWLs that change over time. or is this already a result of your study?

This was not the outcome of this study but a general advantage of the model and references will be provided.

L304: Further explanation needed. The figure does not provide any details, especially no comparison, as written by the authors.

We agree these 2 sentences are confusing. It is also true that the difference between the various models is never extremely noticeable, because all the models performed well eventually. A thorough examination of the results of figure 3 (comparison of the 3 model types in single-station mode) and of figure 5 with figures A1 and A2 led us to the conclusion that GRU performed slightly better. Another reason why GRU was preferred is also related to its computational efficiency. Since the difference in performances is not very noticeable, we suggest the following modification:

"All models tested in the case of this study, performed more or less equivalently and eventually led to very satisfactory results. This can be attested by performance comparison shown in figure 3 (comparison of the 3 model types in single-station mode) and by comparing figure 5 with figures A1 and A2 (multi-station mode). We finally decided to favor the GRU architecture owing to its recognised computational efficiency over more traditional LSTM-based architectures (Cho et al., 2014; Cai et al., 2021; Chidepudi et al., 2023)".

L355: This is an information you expect earlier in the manuscript.

Agreed, we'll move this information to the data section for better context.

L372: Why do you formulate "new research questions" here, is this necessary?

We agree, formulating new research questions again at this stage can be misleading. We then removed them.

L425: No_ohe_no_stat approach?

We'll update to use consistent and clear naming conventions for all approaches throughout the paper.

References: Nourani, V., Alami, M. T., & Vousoughi, F. D. (2015). - I do not find a citation of this paper.

Corrected the citation in line 111

References:

- *Heudorfer, B., Liesch, T., & Broda, S. (2024). On the challenges of global entity-aware deep learning models for groundwater level prediction. Hydrol. Earth Syst. Sci, 28, 525–543. https://doi.org/10.5194/hess-28-525-2024*
- *Haaf, E., Giese, M., Reimann, T., & Barthel, R. (2023). Data-driven estimation of groundwater level time-series at unmonitored sites using comparative regional analysis. Water Resources Research, 59, e2022WR033470. https://doi.org/10.1029/2022WR033470*
- **Slimani, S., Massei, N., Mesquita, J. *et al.* Combined climatic and geological forcings on the spatio-temporal variability of piezometric**

levels in the chalk aquifer of Upper Normandy (France) at pluridecennal scale. *Hydrogeol J* 17, 1823–1832 (2009). https://doi.org/10.1007/s10040-009-0488-1

- El Janyani, S., Dupont, JP., Massei, N. *et al.* Hydrological role of karst in the Chalk aquifer of Upper Normandy, France. *Hydrogeol J* 22, 663–677 (2014). https://doi.org/10.1007/s10040-013-1083-z
- Sanae El Janyani, Nicolas Massei, Jean-Paul Dupont, Matthieu Fournier, Nathalie Dörfliger. Hydrological responses of the chalk aquifer to the regional climatic signal, Journal of Hydrology,Volumes 464–465,2012, Pages 485-493,ISSN 0022-1694, https://doi.org/10.1016/j.jhydrol.2012.07.040
- Giese, M., Haaf, E., Heudorfer, B., & Barthel, R. (2020). Comparative hydrogeology – reference analysis of groundwater dynamics from neighbouring observation wells. *Hydrological Sciences Journal*, *65*(10), 1685–1706. https://doi.org/10.1080/02626667.2020.1762888
- Haaf, E., Giese, M., Heudorfer, B., Stahl, K., & Barthel, R. (2020). Physiographic and climatic controls on regional groundwater dynamics. *Water Resources Research*, 56, e2019WR026545. https://doi.org/10.1029/2019WR026545
- Heudorfer, B., Haaf, E., Stahl, K., & Barthel, R. (2019). Index-based characterization and quantification of groundwater dynamics. *Water Resources Research*, 55, 5575–5592. https://doi.org/10.1029/2018WR024418
- Lee, S., Lee, KK. & Yoon, H. Using artificial neural network models for groundwater level forecasting and assessment of the relative impacts of influencing factors. *Hydrogeol J* 27, 567–579 (2019). https://doi.org/10.1007/s10040-018-1866-3
- Tanja Liesch, Andreas Wunsch, Aquifer responses to long-term climatic periodicities,Journal of Hydrology,Volume 572,2019,Pages 226-242,ISSN 0022-1694,https://doi.org/10.1016/j.jhydrol.2019.02.060
- Hejiang Cai, Haiyun Shi, Suning Liu, Vladan Babovic, Impacts of regional characteristics on improving the accuracy of groundwater level prediction using machine learning: The case of central eastern continental United States,Journal of Hydrology: Regional Studies, Volume 37,2021,100930,ISSN 2214-5818, https://doi.org/10.1016/j.ejrh.2021.100930
- Cho, K., Van Merriënboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*
- Chidepudi, S. K. R., Massei, N., Jardani, A., Henriot, A., Allier, D., & Baulon, L. (2023). A wavelet-assisted deep learning approach for simulating groundwater levels affected by low-frequency variability. Science of The Total Environment, 865, 161035.