Response to Reviewer 1 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 your insightful comments and suggestions for improving our manuscript. Please find our point-by-point responses below in bold.

Chidepudi et al. used deep learning approaches to simulate and predict groundwater level dynamics. Authors compared and discussed the performance of different approaches of different combinations, such as different DL models, different inputs (i.e., dynamic factors and static factors), wavelet decomposition of precipitation, one-hot encoding etc. Using deep learning approach to simulate and predict dynamic groundwater levels is challenging. This work is important and could be a good reference for the community. The paper is generally well organized but there are still a lot of details unclear. Major revision is needed for further review.

Thank you for acknowledging the importance and potential impact of our work. We will address all the unclear details in the revised manuscript as described below.

1. *There are no clear introductions of model structures.*

We will enhance the details in the revised version. Specifically, we will include a new subsection in Section 3.1 (Theoretical modelling background) after line 205, providing details of the model structures, including the number of layers, units, and other relevant other common hyperparameter details (as shown in Table 1) for LSTM, GRU, and BiLSTM models. Though these models are widely used nowadays in all other subfields of hydrology as highlighted in line 224, if needed we will provide cell diagrams in supplementary document to explain the main principle of each network type.

Table 1: Hyperparameter details (Modified and adapted from chidepudi et.,al 2023)

Hyperparameter	Value considered
Sequence length	**48**
Dropout	**0.2**
Optimizer	**ADAM**
Early stopping	**50**

Number of layers	**1**
Hidden neurons	**(10, 20, …,100) by 10**
Learning rate	**(0.001,0.01) (log values)**
Batch size	**(16, 32, …,256) by powers of 2**
Epoch	**(50, 100, …,500)**

2. *I didn’t find details of the model input or the structure of the input data. I especially wanted to know this in the multi-station approach*

We will clarify the input data used for each of the multi-station approaches with standalone and wavelet assisted models by enhancing Section 3.2 (Experimental design) after line 240. This subsection will explain the structure and preprocessing of the input data, including the dynamic variables and static attributes, and how they were combined and formatted as input to the models. All the models use sequences as input for point simulation. The input data is structured as a 3D tensor with dimensions (samples, sequence_length, num_features), where the sequence length is set to 48 (4 years of monthly data), and the number of features includes both dynamic (time series of precipitation, temperature, surface net solar radiation…) and one-hot encoded static variables depending on the type of approach. For wavelet models, dynamic variables are also time series that are wavelet components of original inputs (time series of precipitation, temperature, surface net solar radiation…).

3. *How did you choose the training and test sets?*

In Section 3.2 (Experimental design), after line 295, we will add a new paragraph detailing the selection of training and test sets for the different modeling approaches. For the single-station approach, the data was split into training (80%) and testing sets (20%) as described in Chidepudi et al., 2023. Furthermore, to facilitate hyperparameter tuning, the last 20% of the training data was used as a validation set. For the multi-station approach, the train-test split was also performed at each station, following the same procedure as the single-station approach. However, the data from all stations was then collectively combined

during the training. The rationale behind the specific train-test split is to ensure that the models capture the multi-annual to decadal variability in groundwater levels (GWLs) observed in the region. To achieve this, a minimum of 34 years of data (1970-2014) was used for training, while the most recent 8.66 years of data (2015/01-2023/08) were reserved for testing. This split corresponds to approximately 80% of the data for training and 20% for testing. By following this approach, we aimed to ensure that the models were exposed to a sufficiently long period of data during training, enabling them to capture the amplitude and variability of GWL fluctuations over multi-annual to decadal timescales. The testing period was chosen to be the most recent years, allowing for an evaluation of the models' performance on the latest available data.

4. *I didn't find how large your research area (only a figure). The resolution of ERA5 is low and the true variations of these hydrometeorological variables may not be accurately presented by the products*
5. *What do you think about the uncertainties of data products from ERA5*

Common response for both these comments (4&5) as they seem somehow related.

Regarding the research area, we will include additional details on research area in Section 2 (Study Area) to clearly specify the geographic extent covered in our study which is approximately 80,000 km2 covering two main basins (Seine and Somme).

In Section 2 (Data), after line 160, we will also discuss the implications of spatial resolution on capturing local variations when using data products like ERA5.

While we understand your concern about the potential limitations in accurately representing localized groundwater dynamics, ERA5 is the best available global reanalysis with the data available from 1940 and is generally considered adequate for capturing regional and global hydrometeorological variations. ERA5 Reanalysis data do have the uncertainty related to potential regional biases, this is ongoing debate as being discussed in (Maria Clerc-Schwarzenbach et al., 2024.) CAMELS (ERA5) vs CARAVAN (ERA5-Land) paper. Precipitation is considered to have more bias than temperature. However, for our study area, we have been evaluating different potential alternative reanalysis products, such as the Safran reanalysis developed specifically for France (Vidal et al., 2010). It appeared that both ERA 5 and Safran precipitation contained the same low-frequency components as detected in GWL time series as displayed in Fig.2 (this paper) and Fig.11 in Chidepudi et al 2023. ERA 5 then seems quite suitable for our purpose.

Discussing uncertainty of ERA5 is beyond the scope of this paper and can be considered research work as itself. However, we will add relevant references that discussed this point.

6. *Did you only conduct the wavelet decomposition on precipitation or other variables also?*

We will clarify that wavelet decomposition is done only on input dynamic variables after line 200: wavelet decomposition is being performed on time series only, each input time series being eventually replaced with its 5 wavelet components (corresponding to the decomposition level selected).

7. *What is the resolution of the data products of static attributes?*

In Section 2 (Data), after line 165, we will provide information on the resolution and sources of the static attribute data used

Static attributes are available for different ranges of aquifer classes with different resolutions, and we took the one that was associated with the Well IDs. Static attributes, coming from BDLISA database, are point-scale information, i.e., each well received set of attributes given different possible methods (geographical imputation, rule-based, human expertise). BDLISA is based on a mix of information coming from geological maps at a scale of 25km, piezometric maps, hydrochemistry, etc.

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. This should be understood as a local to regional description of aquifers.

8. *What do you think the effects of hydraulic conductivity, elevation, slope etc. static attributes.*

The decision to include the relevant static attributes comes from a trade-off between transposability of models and availability of attributes, as we have to make sure that all those variables are widely available at required resolution. Also, for some attribute 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. For instance, 25km resolution might not be relevant when aquifers are highly heterogeneous. Exploring the role of static attributes in more details would require much further works than what was conducted in this study.

9. *Location of the well, i.e., in confined or unconfined aquifers may also be important*

All the wells considered in the study are in unconfined aquifers.

We would be happy to respond to any further comments while the discussion phase is still in progress.

References

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. https://doi.org/10.1016/j.scitotenv.2022.161035

Maria Clerc-Schwarzenbach, F., Selleri, G., Neri, M., Toth, E., van Meerveld, I., & Seibert, J. (2024). HESS Opinions: A few camels or a whole caravan? https://doi.org/10.5194/egusphere-2024-864

Vidal, J.P., Martin, E., Franchistéguy, L., Baillon, M., Soubeyroux, J.M., 2010. A 50-year high- resolution atmospheric reanalysis over France with the Safran system. Int. J. Climatol. 30 (11), 1627–1644. https://doi.org/10.1002/joc.2003