Review 1

This manuscript evaluates the performance of two machine learning models in predicting groundwater (GW) levels across a dataset of ~5000 wells in Germany. The study examines the influence of both dynamic and static input features on the accuracy of GW level predictions and seeks to enhance the understanding of hydrogeological systems.

The objectives, methodology, results, and discussion are clear, well-structured, and thoroughly explained. The study aligns well with the scope of the journal, and HESS readers would benefit from and appreciate its findings. In my opinion, the manuscript is close to its final form. However, I would like to raise the following points for consideration:

We thank Reviewer 1 for the positive assessment of our manuscript. Below we address the suggestions (marked in blue).

- The manuscript specifies particular values for hyperparameters (e.g., dropout rate, batch size, etc.). Are these values based on specific rules or conventions? Did you test alternative values? While this may not significantly affect the overall conclusions, I believe it would be helpful to clarify this for the reader.

Thank you for your suggestion. We initially selected hyperparameter (HP) values based on empirical heuristics recommended by domain experts, aiming to reduce overfitting and minimize training time. As noted in line 160 ff., the final values for dropout rate and batch size were chosen to achieve these goals. The decision to set the number of training epochs to ten was based on the convergence of the training and validation loss observed during preliminary testing. In particular, models based on the N-HiTS architecture would converge before the 10th epoch.

We will modify section 2.3 (lines 160 ff.) accordingly:

“All model variants were trained with ten different random seeds to account for the stochasticity in the initialisation of model weights. Large batch sizes were used (TFT: 4096, N-HiTS: 1024) to avoid overfitting and to accelerate the training. The risk of overfitting was further reduced by the application of early stopping on the validation loss, a dropout rate of 0.2, and learning rate scheduling using stochastic weight averaging after the second epoch (Izmailov et al., 2019). Thus, the selected hyperparameter values were based on empirical heuristics recommended by domain experts, aiming to reduce overfitting and minimize training time. All model variants were trained for a maximum of ten epochs, a duration sufficient to ensure model convergence. In many cases, training terminated earlier due to the implementation of the early stopping criteria.”

- Is there a specific reason for setting the prediction horizon to a maximum of 12 weeks?

The aim of our study was to provide seasonal groundwater level predictions. The different seasons are known for their substantial impact on groundwater recharge, and thus on groundwater levels. Accordingly, we selected a prediction horizon of 12 weeks, equivalent to approximately three months, as an appropriate timespan for reflecting seasonal patterns. Furthermore, we observed a decrease in model performance with longer horizons as shown in figure 3. A 12-week prediction horizon allowed us to maintain acceptable predictive performance across the more than 5,000 monitoring wells.

Figures B9 and B10 indicate that attention is higher one year before the prediction than at times closer to it. Could you elaborate on why this happens?

We interpret the results in figures B9 and B10 as likely related to the autocorrelation function of many of the observed groundwater level time series. Based on the intrinsic feature importance of the TFT, we know that the most important feature is the historical groundwater level. For the one-week prediction, the groundwater level from the corresponding week one year prior to the prediction is likely the most influential information for the TFT models, as reflected in the attention scores. This is likely due to the seasonality observed in many groundwater monitoring wells. Accordingly, for the 12-week prediction, the groundwater levels from the week one year prior and 12 weeks before the prediction are the most important time steps. To support our interpretation, we will include the autocorrelation for the 52nd week as well as the 12th week of each groundwater hydrograph in the Supplement as additional explanation. We will also add a reference in the description of figures B9 and B10 to the autocorrelation.

- In figure B10, why are attention values not zero in the interval of 0–10 weeks? Does this imply that the algorithm is somehow using inputs from these time steps? I suggest including a diagram to illustrate how inputs and outputs operate in the ML algorithms (e.g., similar to Figure 1 in Kratzert et al., 2018). This would help clarify which specific information is being utilized and when.

We thank the Reviewer for the suggestion to clarify how input features operate in the models. We will add a figure in the method section.
Both architectures follow an encoder-decoder structure. The encoder creates a latent representation of the input features, while the decoder uses this representation to generate predictions. The historical groundwater levels, historical climatic features and the static features are processed in the encoder. In the decoder, the climatic features for the prediction horizon (so-called future knowns) and the static features are used, while groundwater levels are not used. The non-zero attention values observed during the 12-week prediction likely reflect the information the Temporal Fusion Transformer (TFT) extracts from the future known features during these time steps.