Review 3

The submission by Kunz et al. presents the development and application of a machine learning model for groundwater level prediction in Germany. The models are referred to as "global" since they are trained against a multitude of wells simultaneously. The study entails several novel aspects which make the submission highly relevant for publication in HESS: 1) the applied models have not previously been applied in the groundwater domain and go beyond the state of the art, 2) such a large number of monitoring wells with time series data has not been used for model development before, and 3) the thorough investigation of the effect of static features in the models.

We thank Reviewer 3 for the positive assessment of our manuscript. Below we address the suggestions (marked in blue).

I only have a few comments that I wish to see addressed before publication:

1. **Introduction**: The cited literature in the introduction could be diversified. Here are two suggested references that could be included:
    - Collenteur, R. A., Haaf, E., Bakker, M., Liesch, T., Wunsch, A., Soonthornrangsan, J., ... & Meysami, R. (2024). Data-driven modelling of hydraulic-head time series: results and lessons learned from the 2022 Groundwater Time Series Modelling Challenge. Hydrology and Earth System Sciences, 28(23), 5193-5208.
    - Chidepudi, S. K. R., Massei, N., Jardani, A., & Henriot, A. (2024). Groundwater level reconstruction using long-term climate reanalysis data and deep neural networks. Journal of Hydrology: Regional Studies, 51, 101632

Thank you for the suggested literature. We will add the references in the introduction where studies for groundwater level prediction are discussed (line 4 ff.).

2. **Section 2.1.1**: This section is missing information on the temporal resolution of the data. What is the frequency of the measurements, and were the measurements aggregated in time?

The temporal resolution of the groundwater measurements was weekly after the data preprocessing steps. This resolution of the groundwater levels is reported in table 1. Prior to the preprocessing, some of the raw groundwater level observations were in part on monthly resolution. Those groundwater level observations have been upsampled via linear interpolation before we obtained the data.

3. **Section 2.3**: Please clarify if the models are run in an autoregressive manner, simulating one timestep at a time (i.e., prediction at t1 is added to the dynamic inputs to predict t2), or if a sequence for the entire forecast horizon is outputted directly.

Thank you for the suggestion. The models generate predictions for entire sequences at once, i.e. seq2seq prediction. In our case, the model predicts a sequence of groundwater level values for 12 weeks. We will clarify this in section 2.3 (line 159 ff.):

“Both ML architectures used in this study are designed for sequence-to-sequence predictions. During training, the models processed an input sequence autoregressively and predicted an output sequence of groundwater levels. For each time step, a look-back window (i.e. sequence length) of 52 weeks for the dynamic features was used to represent one annual cycle. Groundwater levels were predicted for 12 weeks. During the 12 week prediction the model has access to the exogenous dynamic features, but not to the groundwater level.”

4. **Section 2.4**: Please clarify how the prediction intervals have been utilized. Were three separate models trained for the 0.1, 0.5, and 0.9 quantiles?

The prediction intervals were obtained with one model. The model learns to predict different parts of the conditional distribution simultaneously. Computing the loss for multiple quantiles results in a multi-output model, with output dimensions (horizon x quantiles). During training the loss is reduced via averaging.

5. **Discussion**: Given the data presented in this paper, I was hoping the authors would attempt predictions at ungauged wells. Currently, groundwater level observations are used both in the dynamic and static features, making predictions at ungauged wells impossible with the existing model setup. I encourage the authors to add a discussion section outlining a path towards predicting groundwater levels at ungauged wells. This could be supported with an additional model experiment that excludes observed groundwater level data from the input features and is based on a spatial hold-out of monitoring wells for model testing. Even a poor test performance of such a spatio-temporal holdout experiment would be relevant to publish to underline the need for future research. To my knowledge such an experiment has not been published yet

For seasonal or short term prediction all available information, i.e. both endogenous (lag features or historical measurements) and exogenous inputs, are typically used to improve the predictive performance. Thereby, the historical measurements of the target feature serve as a starting point for the prediction. For long-term predictions such as decadal predictions or when the aim is to predict wells a strategy without the target feature as input feature (exogenous-only) is necessary, to which the reviewer is referring to. However, this was not the aim in our study.

Nevertheless, in preliminary experiments, we have evaluated the importance of historical groundwater levels as input by making predictions with the Temporal Fusion Transformer (TFT) without the historical groundwater level as an input. The model's performance deteriorated with an estimated NSE of 0.37 (with static features) and 0.22 (purely dynamic) for the one-week prediction, and an NSE of 0.08 (with static features) and -0.04 (purely dynamic) for the 12 week prediction. We will include these analyses in the supplement and mention the results in the manuscript. As the performance of the TFT model across all 5,288 monitoring wells was considerably worse without the groundwater levels as input feature, we decided not to pursue further analyses in this regard. It is important to note that our study includes a large number of monitoring wells with varying degrees of predictability, including many located in hydrogeological complex areas or influenced by anthropogenic effects. This contrasts with other studies that often focus on wells with inherently high predictive capacity (e.g. wells predominantly influenced by climatic factors), leading to higher NSE values. Nevertheless, there are still 884 wells in our study where the TFT model provided with static features achieved an NSE >= 0.5 even without groundwater levels as an input feature.

Moreover, we believe that the topic of prediction in ungauged wells is beyond the scope of this study. The main focus of our study is on seasonal predictions, where the inclusion of historical measurements of the target feature as input is standard practice to enhance the models performance. Removing this critical feature would not align with our study's objectives, which aim to evaluate the utility of advanced machine learning architectures in a realistic operational setting. Heudorfer et al. (2024) have carried out analyses on the predictive capabilities on ungauged wells with a global LSTM model.

References

Heudorfer, B., Liesch, T., & Broda, S. (2024). On the challenges of global entity-aware deep learning models for groundwater level prediction. *Hydrology and Earth System Sciences*, *28*(3), 525–543. https://doi.org/10.5194/hess-28-525-2024