Overview

The authors develop an entity-aware deep learning model for spatially and temporally continuous groundwater level prediction using a combined Long Short-Term Memory (LSTM) and MultiLayer Perceptron (MLP) network. They rely on ground observations from 108 wells in Germany and other dynamic and static predictor data obtained from multiple sources to train, validate, and test the model. The authors also perform some interesting comparisons of four model variants, namely, the time series feature-driven model (TSFeat), environmental feature-driven model (ENVfeat), random static features (RNDfeat), and dynamic inputs only features (DYNonlyfeat). While there are some issues with spatial generalizability in the out-of-sample setting, the model shows satisfactory performance with the Nash-Sutcliffe Efficiency (NSE) > 0.8 in an in-sample setting.

Overall, the manuscript is generally well-structured, with detailed explanations of the methodologies and data. However, the following comments should be addressed before this manuscript is published.

> We thank Reviewer 2 for his positive assessment. We will address his valuable suggestions as itemized below in our responses (blue). For the sake of a clear overview, all references initially provided by Reviewer 2 were omitted from this answer. They can still be found in the original Review.

Major Comments

The authors should discuss relevant literature that incorporates process-based, machine learning, or hybrid models and remote sensing data for groundwater level monitoring. The Introduction section should highlight the relevance of this topic more and refer to some of the negative impacts of groundwater depletion and why groundwater level monitoring is essential.

> Reviewer 2 is right in that the introduction could use some further expansion on the need for groundwater level monitoring and modeling due to threats to sustainable groundwater use as lined out beautifully by the literature provided by Reviewer 2. We will add some more sentences on these aspects in a revised version.
>
> Regarding discussion on process-based models, we refer to line 21-26, where it is argued that numerical models are not suitable for the national scale. Accordingly, there is no national numerical groundwater model that could be discussed. We will add this fact to the discussion in line 21-26, and will expand on alternative applications like e.g. the process-based estimation methods in the central valley study provided by Reviewer 2 – which are, however, also unavailable in Germany.
>
> Regarding remote sensing data studies, we point out that our study focusses primarily on analyzing a general theoretical problem with entity aware models, leaving the question open how remote sensing data could be generally beneficial in this.

In addition to the predictor data summary, the authors should include a description of the predictor data listed in Tables 1 and 2 and the associated uncertainty. In Table 2, what does selfderived mean? Would snow water equivalent and soil moisture be helpful as additional predictors to capture the groundwater dynamics better? The authors should make a stronger case for selecting HYRAS 3.0 than other globally available land-surface models like the Global Land Data Assimilation System (GLDAS), which provides spatially and temporally continuous estimates of various hydrological processes acting as critical drivers of groundwater dynamics.

> We agree with Reviewer 2 that descriptions of static features in table 2 are rather brief for readers who are unfamiliar with this data. We will add extended descriptions of the static features and the associated uncertainty in the revised manuscript.
>
> In table 2, self-derived means that we did not take these static features from an existing dataset, but we calculated them from the dynamic meteorological input features ourselves. We will clarify this in the table, thank you for highlighting the ambiguity.
>
> Yes, additional dynamic input features such as snow water equivalent, soil moisture or others have the potential to positively impact model performance. However, the overall best performance was not the scope of this study, which deals with theoretical considerations regarding entity awareness.
>
> We relied on the HYRAS dataset, as it has proved its suitability in several studies before (e.g. Wunsch et al. 2021, 2022), and has a higher spatial resolution than the global datasets available. Moreover, we used the same meteorological dataset as in Wunsch et al. (2022) allowing a better comparison with their results for the single well method.
>
> We will sharpen the formulation of the research aims in the introduction, and the data section regarding HYRAS in order to make this point clearer.

The authors should include the model forecasts beyond January 2016. While it may be challenging to obtain in-situ groundwater levels between 2016-present, it would be interesting to observe how the model predictions compare to the GRACE- and GRACE Follow-On (GRACEFO)-based total water storage changes (https://grace.jpl.nasa.gov/data/data-analysis-tool/) at a regional or national scale. This comparison would serve as an additional model validation and strengthen the manuscript.

> We agree that inclusion of model forecasts beyond 2016 would be advantageous. However, because we rely on a previously published dataset, we have no way of updating this data.
>
> Also, we argue that GRACE could not feasibly be used as a substitute for groundwater level measurements or as a comparison within the scope of our study, due to its inherent coarse spatial and temporal (monthly) resolution. Moreover, it constitutes a different variable (total water storage changes, as opposed to groundwater level directly) with inherent uncertainties in the computation of groundwater storage and subsequently groundwater levels, relying on various additional data. This would distort the original scope of our study, which focuses on theoretical and methodological considerations regarding entity awareness.

There should be an additional section (or a subsection within the Introduction) describing the study area and related studies on groundwater level changes. Also, the spatial distribution of the 108 well locations should be shown on a map.

> We agree with Reviewer 2, that a (sub-)section describing study area, accompanied by a map, will be useful. We will include it.
>
> Regarding a (sub-)section on ‘related studies on groundwater level changes’, it is unclear what reviewer 2 means by this broad formulation. There are multiple entire research fields occupied with groundwater level changes. We are confident that we reviewed the literature relevant to our study’s domain allocation appropriately in the introduction, but will happily update this discussion with the resourced provided by Reviewer 2 (e.g. the ones provided in Reviewer 2’s final major comment).

What are the 11 land cover classes in the CLC data? How are these used in the model? Can the categories be reduced by aggregating to a base class? E.g., crops aggregated to ‘Agriculture,’ urban/industry to ‘Urban,’ and so on? Is there no significant change in built-up or irrigated areas within the temporal domain of the model? The potential effects of land use changes on the model performance should be discussed. Also, the percentage of land use classes should be described in the Study Area section.

> The CLC is used as a one-hot-encoded static feature input to the ENVfeat model variant, next to the other 17 static environmental features. Yes, they could be reduced to fewer classes, there are e.g. 3 forest classes that could be combined into one, and 4 different urban classes that could be combined. However, from our conceptional understanding, using the single classes as defined makes more sense (e.g. the groundwater recharge in coniferous forest is significantly smaller than in deciduous forest, thus groundwater level should react differently in both forest classes to meteorological inputs, the same applies to continuous and discontinuous urban fabric etc.).
>
> Second, yes, there can be land use change over time, but in general, Germany ‐ as a highly developed and densely populated country where use of land is subjugated to densely layered interests with extensive laws preventing unauthorized land use change ‐ has very limited land use change over time. There is one notable exception, namely that about 8% of the countries area switched from arable land to forests over the period 1982-2016 (Song, 2018). All other land use types remained stable. We therefore consider land-use as a quasi-static feature. We will add text to explaining this aspect in the revised version of the manuscript.

The corresponding time series of the dynamic predictors for the two wells in Figure 5 should be added and tied up with the discussion related to the permutation feature importance.

> Thank you for this suggestion. We could add a figure that includes the dynamic predictors (at least P, T and rH) for the shown wells. However, we think that this would overload figure 5 at the present state and distract the reader from the actual point of discussion in figure 5 (i.e. the comparison of the in-sample and out-of-sample performance). If you insist, we suggest adding a separate figure. However, we think that setting the dynamic inputs features for selected single wells into relation to the permutation feature importance, which shows an average importance over all wells, is difficult anyway.

Evapotranspiration (ET) is the second largest component of the water cycle after precipitation (https://openetdata.org/what-is-evapotranspiration) and is a critical driver of groundwater use, which, in turn, is correlated to groundwater levels (Majumdar et al., 2020; 2022; Brookfield et al., 2023; Melton et al., 2021; Senay et al., 2022). Why didn’t the authors include it as a dynamic predictor and instead rely on the potential ET (Table 2)? While the OpenET and the Landsat-derived actual ET products are currently available only over the conterminous United States (CONUS), the globally available 500 m MOD16 actual ET is available within the temporal domain of the model. Thus, the authors should justify the choice of their predictors.

> We agree that ET is an important dynamic predictor for groundwater levels, which could potentially improve the overall model performance. But as pointed out already above, we stick to the HYRAS dataset for the dynamic meteorological inputs (which does not include ET as a modelled parameter) for several good reasons. Moreover, ET is mainly controlled by temperature and relative humidity, which are included in our dynamic predictors.

Lines 50-60: While the proposed machine learning-based method of using multiple wells to develop an entity-aware global groundwater level prediction model is new, earlier studies have integrated remote sensing, climate, and hydrogeologic data in a machine learning framework for estimating annual groundwater withdrawals (Majumdar et al., 2020; 2021; 2022; Wei et al., 2022) and land subsidence (Smith & Majumdar, 2020; Hasan et al., 2023). For the studies on groundwater withdrawal estimation, a single machine learning model was trained and validated using in-situ pumping measurements from multiple wells across vast geographical areas (states of Kansas and Arizona in the U.S.). Thus, the authors should clearly convey that the novelty lies in groundwater level monitoring rather than the entire hydrogeology domain.

> We thank Reviewer 2 for highlighting this research. We were not aware of these studies and will include them in the introduction section where existing global models are discussed (line 50 ff.), and point to the novelty concerning groundwater level modelling.

Minor Comments

Line 94- Fix typo: followed by a brief *introduction.*

> We will fix this, thank you.

What are the spatial resolutions of the predictor data listed in Table 2? How do the authors map the groundwater wells to these gridded raster datasets?

> Thanks for spotting this. We forgot to add the information that environmental feature values were simply selected at the location of the respective groundwater well. We will add this information and will specify resolution in Table 2.

Report other error metrics like the coefficient of determination (R2), root mean square error (RMSE), and the mean absolute error (MAE) in Table 3.

> Yes, we can report other error metrics. We will add R^2 and KGE in table 3 (Reviewer 2 probably means table 3 here). RMSE and MAE are not suitable, because they are not comparable between different sites due to different reference height, amplitude etc.

The CLC acronym is not defined.

> Thank you for pointing this out. CLC first appears in Table 2, we will add definition there.

Do the authors scale all the features? What scaling is applied?

> As mentioned in line 187, all dynamic features were standardized, i.e. standard scaled. However, we realized that the scaling method for the static features is not mentioned in the paper. Thank you for pointing this out, we will add this information in section 2.2 or 3.1. Numerical static features were standard scaled as well, categoric static features were one-hot encoded.

Lines 235-240: For the out-of-sample setting, are the scores only calculated for the testing period of a well that has been left out of model training? Why not calculate the score for the entire period?

> Yes indeed, scores are only calculated for the testing period of each well. We will point this out in the revised version. From the aspect of data leakage, it would have been possible to include the scores for the entire period of the left-out wells in the out-of-sample setting. However, we decided to stick with the same test period to allow direct comparability with the scores in the in-sample setting. As reviewer #1has correctly remarked, there is a (for groundwater level time series practically unavoidable) bias in the test set, thus, changing it would mean to lose this comparability.

References:

Song, X.P., Hansen, M.C., Stehman, S.V., Potapov, P.V., Tyukavina, A., Vermote, E.F. and Townshend, J.R., (2018). Global land change from 1982 to 2016. Nature, 560(7720), pp.639-643. https://doi.org/10.1038/s41586-018-0411-9

Wunsch, A., Liesch, T., & Broda, S. (2021). Groundwater level forecasting with artificial neural networks: a comparison of long short-term memory (LSTM), convolutional neural networks (CNNs), and non-linear autoregressive networks with exogenous input (NARX). Hydrology and Earth System Sciences, 25(3), 1671-1687, https://doi.org/10.5194/hess-25-1671-2021

Wunsch, A., Liesch, T. and Broda, S. (2022) 'Deep learning shows declining groundwater levels in Germany until 2100 due to climate change', *Nature Communications*, 13(1), p. 1221. Available at: https://doi.org/10.1038/s41467-022-28770-2