The authors have presented an application of machine learning techniques to create a global model of groundwater levels in Germany. They compared two different input model settings: one with static features and one with dynamic features. Additionally, they compared these settings with two reference cases, one with random entity variables and a second without entity variables. Their results indicate that both model settings perform well under in-sample conditions, but their performance diminishes in out-of-sample conditions. The work aligns with the growing trend in this field, introducing entity-aware methods to hydrogeology and yielding promising results. I believe the paper has the potential for publication, but there are some points that need to be addressed before publication.

> We thank Reviewer 2 for his positive assessment. We will address his valuable suggestions as itemized below in our responses (blue).

Limited Machine Learning Methods Tested: The authors only tested two machine learning methods, LSTM and CNN. There are powerful alternatives like Transformers, which have outperformed LSTM in other studies (https://doi.org/10.1016/j.apr.2023.101833). LSTM models are designed for handling transient conditions, whereas CNNs are adapted to do so. There are various other methods like extreme gradient boosting that have been applied in hydrogeology that can be a powerful alternative to CNN (cite: https://doi.org/10.1016/j.watres.2023.119745, https://doi.org/10.1007/s10661-020-08695-3, https://doi.org/10.1016/j.scitotenv.2021.151065). The authors should explain why they chose not to explore more advanced models before attributing out-of-sampling prediction issues to dataset limitations.

> Reviewer 1 is right in that potentially more powerful AI models might be available, when considering absolute performance. However, there are sound reasons why we chose to stick with LSTM (and CNN) models. First of all, as outlined in the introduction, and as elaborated in detail in a recent review by Tao *et al.* (2022) cited in the paper, LSTM and CNN are currently the predominant model class in groundwater level modelling. This is because they consistently deliver high-class performance while maintaining some degree of model simplicity, thereby satisfying Occam's Razor. The mentioned extreme gradient boosting proved to be powerful in groundwater quality modeling. However, this is a related but different field from groundwater level modeling with a number of significant differences, making methods not directly transferable. Transformers seemed promising 1 or 2 years ago, and their suitability as a general-purpose method beyond its original domain (language modeling) are increasingly called into question because they can be outperformed by more simple linear DL models (Zeng *et al.*, 2023), notably across the board (Das *et al.*, 2023). In conclusion, while testing different newer/other models to increase overall performance is desirable, this needs a separate study with careful model selection and a clear experimental design geared towards appropriate research questions. This was out of scope for the study at hand, which deals with theoretical questions regarding entity awareness. Regarding this manuscript, we will further elaborate the choice of LSTM and CNN over other methods in the papers' introduction along the line of above argumentation.

Data Fusion Method Comparison: In Line 164, the authors propose a new data fusion method. Did they compare this method to existing methods to demonstrate its benefits and limitations? Comparing it to cited methods would provide valuable insights.

> The data fusion method used in the study is not new but well-established in various fields (e.g. Liu et al. 2022, Miebs et al., 2020). However it is true that it in other studies in neighboring disciplines (e.g. Kratzert et al. 2019), often a more simplistic approach of duplicating the static input features in each time step is used. As discussed e.g. in Miebs et al.

> (2020), this is not an optimal choice for an RNN architecture. Such an approach leads to a significant increase in the number of RNN parameters, since duplicated static features are evaluated each time for every sequence. Moreover, these duplicate data do not add any meaningful additional information. As a consequence, the training of such a network is both memory- and time-consuming (in our case it tripled the computation time in some initial experiments and, moreover, showed rather similar results). We are also aware that there are more sophisticated approaches of combining dynamic and static features in machine learning models (like using static features to initialize cell states in LSTMs or to learn attention weights). But even though we agree with the reviewer, that a comparison of different data fusion methods would provide valuable insights, this was not the scope of our study. Again, we chose a well-established method that yielded good results as a basis for our actual question of research.

Introduction: The introduction lacks a description of the study area, which should be addressed since the model applies to a single case study. Adding a figure depicting the study area and well distribution would enhance the paper's context.

> We thank Reviewer 1 for the useful suggestion. We will add a map with study are to the revised manuscript.

Well data selection (Line 104): Quantitatively explain what "spatial coverage as representative as possible" means. Clarify if there is a minimum distance between wells, data density per area, or any specific criteria used for well selection. Provide the original dataset size from which data was picked.

> The groundwater data used in this study was primarily chosen because it is a readily available dataset that is already published (Wunsch et al., 2022), enabling reproducibility and circumventing the need to assemble and publish a new dataset, which is a very time consuming and painful process in Germany due to data accessibility problems, unnecessarily delaying paper publication in a competitive field. We will remove the note on data distribution and will instead expand the justification to use a pre-published dataset for reproducibility and refer the reader to the published paper of Wunsch et al. (2022). We hope this, together with the visual evidence of the datasets' actual distribution from the map (see comment above) will suffice Reviewer 1.

Upscaling (Line 136): Elaborate on the importance of not having too fine-grained categorical data. Describe the upscaling process and how the authors ensured that each category is correctly represented. Provide references or explore the effect of upscaling on training and prediction.

> We realize the passage in the manuscript is written ambiguously. There was no upscaling involved on our side. Instead, we simple had several categorical datasets of the same type at hand and chose the one with less categories. As an example, for soil type, we had the choice between a product called “buek200” with a scale of 1:200,000 which has more than 550 different soil type categories, and another product called “buek5000” with a scale of 1:5,000,000 which has only 23 categories, and which is a generalized version of buek200 on a larger scale. Buek5000 was chosen because communality of soil type classes between groundwater locations would be impossible with buek200: Using buek200 would almost certainly lead to every location having a unique soil type, thus not allowing any study of entity awareness. This is the selection process we wanted to describe in line 136. We will elaborate this better in the text.

MLP Classifier (Line 168): Explain the advantages of adding an MLP classifier rather than providing static features directly. Address concerns about uncertainty propagation due to MLP output in the concatenation.

> We are not sure what reviewer #1 means here. We did not use an MLP classifier, but an MLP for processing the static features (as a regression, not classification). We refer to the point above on data fusion. The chosen approach is a well-established method for the incorporation of static features into recurrent neural networks, see e.g. Miebs et al. 2020.

MLP Output Nodes (Line 170): Specify the number of output nodes in the MLP.

> All numbers of output nodes are given in the text. In line 170 we write "with one fully connected (Dense) layer of size 128 in the static model thread."

Validation MSE (Figure 2): Explain the phenomenon where the validation MSE is smaller than the training MSE, especially in the initial epochs. This could indicate a bias in the validation dataset, and clarification is needed.

> We thank Reviewer 1 to point us towards this aspect that is not sufficiently elaborated yet. The shape of the losscurves indeed indicates bias in validation data, which is probably also the case for the test data. This is due to the fact that validation data is not uniformly sampled over the whole time period, but the fixed time period of 2008-2011 (and test data being 2012-2015), as specified in the paper. These periods are the most recent data period, and it was consciously chosen as fixed due to the fact that the aim is forward prediction of groundwater levels, meaning that the most recent groundwater levels are most representative for a possible future (as opposed of choosing rolling time periods for validation and testing). We will elaborate this better in the revised version of the manuscript.

Feature Importance (Line 245): Suggest using the SHAP method (Lundeberg and Lee, 2017, https://doi.org/10.48550/arXiv.1705.07874) for more stable feature ranking, as it has been used effectively in similar studies (Ransom et al., 2020, https://doi.org/10.1016/j.scitotenv.2021.151065).

> Thank you for this suggestion. We are aware of the SHAP method and we have also used it ourselves before in other studies (i.e. Wunsch et al. 2022). However, it is quite computationally intensive, and that is why we preferred Permutation Feature Importance here. We will better explain the choice of Permutation Feature Importance over other XAI methods (like SHAP or Layerwise Relevance Propagation) in the revised version and we will take the suggestion up in the manuscripts' discussion as a potential alternative method.

Literature

Das, A. *et al.* (2023) 'Long-term Forecasting with TiDE: Time-series Dense Encoder'. arXiv. Available at: http://arxiv.org/abs/2304.08424 (Accessed: 13 November 2023).

Kratzert, F. *et al.* (2019) 'Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets', *Hydrology and Earth System Sciences*, 23(12), pp. 5089–5110. Available at: https://doi.org/10.5194/hess-23-5089-2019.

Liu, Q., Yang, M., Mohammadi, K., Song, D., Bi, J., & Wang, G. (2022). Machine Learning Crop Yield Models Based on Meteorological Features and Comparison with a Process-Based Model. Artificial Intelligence for the Earth Systems, 1(4), e220002.

Miebs, G., Mochol-Grzelak, M., Karaszewski, A., & Bachorz, R. A. (2020). Efficient strategies of static features incorporation into the recurrent neural network. Neural Processing Letters, 51(3), 2301-2316.

Tao, H. *et al.* (2022) 'Groundwater level prediction using machine learning models: A comprehensive review', *Neurocomputing*, 489, pp. 271–308. Available at: https://doi.org/10.1016/j.neucom.2022.03.014.

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.

Zeng, A. *et al.* (2023) 'Are Transformers Effective for Time Series Forecasting?', *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(9), pp. 11121–11128. Available at: https://doi.org/10.1609/aaai.v37i9.26317.