Review 2

I Cannot find the importance of the presents study and how it contributes to the improvement of our knowledge in terms of GWL prediction using machine learning. Predicting GWL using one deep learning model (the CNN) is not new and the fact that the authors propose a modelling strategy based only on exogenous variables is no as important to be proposed and presented as an innovative approach. Furthermore, the fact that three different evaluation strategy, i.e., blocked cross-validation (5 bl-CV), repeated out-of-sample validation (repOOS), and out-of-sample validation (OOS) are compared is not a solid argument to justify the importance and novelty of the present paper. Yet, a modelling strategy based only on one ML model is extremely unsound as there is no any baseline of comparison. The adoption of weekly data is not justified and for closing, section results is extremely poor and unsound. There is no any interpretability of the model and a ranking of the features based on their contribution to the final model response.

Despite the harsh criticism, which we believe is neither substantiated nor justified, we would like to thank the reviewer for their thorough review of our manuscript. We regret that the significance of our study was not fully apparent to them, and we will use this opportunity to provide further clarification in the revised version of the manuscript.

Nevertheless, we would like to clarify a few key points below.

- First, we would like to emphasize that our study did not aim to present the use of exogenous variables or a CNN for prediction as innovative approaches, nor did it aim to compare multiple model architectures. In fact, there is already a wealth of research on these aspects. Similarly, aspects such as model interpretability and feature ranking were not emphasized, as the methodological focus was on the performance evaluation of the validation strategies.

- We aimed to investigate how different validation strategies — blocked cross-validation (bl-CV), repeated out-of-sample validation (repOOS) and classic out-of-sample validation (OOS) — influence the accuracy of performance evaluation. We are very sorry that Reviewer 2 does not consider this sufficient to justify the importance and novelty of the present paper. Various publications in other disciplines have recognized that validation strategies do indeed have a significant influence (e.g. Bergmeir et al., 2014, 2018; Bergmeir & Benitez, 2011; Bergmeir & Benítez, 2012; Cerqueira et al., 2020). We therefore consider it legitimate to investigate whether and to what extent these findings can be transferred to the field of groundwater level prediction. The main relevance of our work lies in our critical analysis of the effects of data splitting strategies and the selection of representative time periods for training, validation and testing. This methodological issue is of great practical and scientific importance, particularly in the prediction of groundwater levels, where time series are frequently non-stationary and exhibit long-term trends or seasonal effects. Incorrectly selecting training or test periods can lead to misleading model evaluations, an issue that is often overlooked in many machine learning applications.

- Regarding the criticism of using a single model (1D-CNN), we would like to clarify that our focus is on comparing validation strategies, not innovating models. This is why an established model architecture was chosen. As we explained in our response to Reviewer 1, the findings on the robustness and transferability of the evaluation methods can be applied to other ML approaches. However, to strengthen the results further, we applied validation methods to an LSTM model. We will include these results, which hardly differ from those of the 1D-CNN, in the revised version of the manuscript.

- Weekly data is commonly used in groundwater level prediction as it accurately represents the dynamics of most aquifers and strikes a good balance between data availability and the volume of training data required. We will take this into account in the revised version of the manuscript.

We hope that these explanations clarify the significance and relevance of our study. Our work provides practical guidance regarding the selection of training, validation and test data periods in ML models for groundwater level forecasting. Our work expands on existing research approaches, which have primarily focused on autoregressive models to date.

References

Bergmeir, C., & Benitez, J. M. (2011). Forecaster performance evaluation with cross-validation and variants. *2011 11th International Conference on Intelligent Systems Design and Applications*, 849–854. https://doi.org/10.1109/ISDA.2011.6121763

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, *191*, 192–213. https://doi.org/10.1016/j.ins.2011.12.028

Bergmeir, C., Costantini, M., & Benítez, J. M. (2014). On the usefulness of cross-validation for directional forecast evaluation. *Computational Statistics & Data Analysis*, *76*, 132–143. https://doi.org/10.1016/j.csda.2014.02.001

Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, *120*, 70–83. https://doi.org/10.1016/j.csda.2017.11.003

Cerqueira, V., Torgo, L., & Mozetič, I. (2020). Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning*, *109*(11), 1997–2028. https://doi.org/10.1007/s10994-020-05910-7