Review 3

This paper reviews and analyzes the validation strategies of time series deep learning models. They classify the metric performance assessment approaches as three groups: out-of-sample validation (OOS), blocked cross-validation (bl-CV), and repeated out-of-sample validation (repOOS). Subsequently, they establish one Dimension Convolutional Neural Networks (1D-CNN) model considering exogenous meteorological inputs with time lags for groundwater level (GWL) prediction. And then, a data set of 100 GWL time series (including 50 stationary and 50 nonstationary time series) in Brandenburg, Germany, are used to assess the validation strategies. Finally, they confirm that bl-CV and repOOS provide the most representative performance estimates for stationary and nonstationary GWL data, respectively. This paper more likes a review of validation strategies, and lacks deep analysis, especially for hydrogeological conditions.

Thank you for your constructive review of our manuscript and your valuable comments. We would now like to address each point individually:

1. Please clarify the contribution of this paper to hydrology.

The focus of our study is to systematically investigate the accuracy of the performance evaluation of validation strategies for machine learning (ML) models. We aimed to stimulate critical examination of the selection of data splits and time periods for training, validation and test data, particularly in the case of non-stationary time series, for which our results are highly relevant. Our work builds on existing studies in other fields that have focused primarily on autoregressive models by examining a non-autoregressive approach and demonstrating how different validation strategies influence performance estimation. We think that these considerations are especially important in hydrology (but also in other geoscientific disciplines) because of the often non-stationary nature of the hydrological (or in general environmental) time series. We will further clarify this in the revised version of the manuscript.

2. Did you consider alteration of the loss functions in the training period to identify suitable hyper parameters for improving the model performance?

To focus methodologically on comparing the validation strategies, we deliberately refrained from further hyperparameter optimization, including adjustment of the loss function. Only the number of training epochs was adjusted to ensure stable training without overfitting. Further investigation of comprehensive hyperparameter optimization could be considered in future work. However, it should be noted that hyperparameter optimization naturally depends on model performance during the selected validation period, which is influenced by the choice of this period. This is a problem in which everything is interrelated. This made it even more important to consider the validation strategies in isolation first, keeping all other parameters constant as far as possible.

3. The better actual model performance of GWL should not only have the small APAE, but also reflect the heterogeneity of aquifers (e.g., response time of GWL to meteorological factors, and amplitude).

We agree that factors such as response times to meteorological influences and the amplitude of groundwater levels are important for interpreting the performance of the model. However, the focus of this study was on the comparative evaluation of validation strategies, which is why standardized performance metrics such as APAE were used.

4. A systematical analysis the hydrogeological conditions of study area are needed (e.g., how many layers of aquifers, which layers the 100 GWL wells located, and groundwater pumping rates), which can help us figure out the best performance model.

A detailed analysis of hydrogeological parameters (e.g. the number of aquifers, the locations of wells and the rates of groundwater extraction) could further enrich the interpretation of model performance. However, the methodological focus of this work was not on achieving the best performance model, but to compare validation strategies using a representative sample of 100 random stationary and non-stationary time series.

The manuscript was submitted to Geoscientific Model Development rather than a purely hydro(geo)logical journal since our study is methodological in nature, examining the performance evaluation of time series models — a topic that may be relevant to, and transferable across, other geoscientific disciplines beyond hydrogeology. Therefore, the insights gained regarding the selection of training, validation and test data splits, as well as robust performance evaluation, are relevant not only for groundwater hydrology, but also for time series-based machine learning (ML) models in geoscientific applications.

In summary, this study makes an important methodological contribution by providing practical guidance on how to reliably evaluate the performance of machine learning (ML) models for time series data, particularly non-stationary time series. The study highlights the need to critically examine the selected validation periods. It shows that the length and location of the validation period within the time series both play a role in correctly assessing performance. This explains why bl-CV and repOOS outperformed the common OOS validation in our study. Thus, this work contributes to a more critical and reflective approach to validation strategies, complementing existing studies that have primarily focused on autoregressive models rather than striving for the best possible predictive accuracy and interpretability. We will further clarify this in the revised version of the manuscript.