Dear **Referee #3,**

We thank you for carefully re-revise our research and suggesting publication. Please find the answers to your final comments below.

**Comments**

(1) *Baseline model is presented in Figure 6 but is not reflected in the text. It needs a small part in the methods and the results/discussion to be fully integrated.*

Thank you for noticing this. We integrated this explanation on the **methods (3.2. Modelling)**: "*For comparison, we employed a baseline model consisting of a sinusoidal function added to the precipitation trend from the last 9 months. This baseline model was optimized using the same Bayesian optimization method, maximizing the NSE and R2 metrics.*" **results (4.1. Modelling)**: *"The baseline model capture the general pattern of GWL fluctuations where the CNN performs higher, but it fails to capture smaller variations."* and on the discussion ""

(2) *L. 127ff The procedure on selecting wells without anthropogenic impact is still not very clear to me unfortunately. The given reference Wriedt et al. (2020) is not very clear in that point also. If possible please provide more detailed information as this is a really important issue that many studies face and others could learn from your procedure. Did you manually check each time series? What metadata is needed to judge this? E.g. Wriedt et al 2020 wrote that "most of the excluded wells were within deeper aquifers".*

Thanks for your comment. Unfortunately, we did not inspect the time series before applying the initial filter by Wriedt et al. (2020), so we are unable to explain that aspect further. However, after applying the filter, we still observed several time series with clear anomalies. We decided to keep them, as they are useful for understanding how different features relate to model performance in our study.

(3) *L. 135ff Thank you for the expansion on the gap filling procedure. One point remains unclear: Euclidean distance of time series … ? Please add between which variables the distance is calculated as this makes a big difference, e.g. "absolute GWL" or somehow transformed GWL? If standardized GWL, how were they standardized?*

We apologize for not clarifying earlier on the text. The Euclidean distance is calculated between GWL time series after standardizing each series to have a mean of 0 and a standard deviation of 1. We then remove any linear trends, allowing us to focus the comparison on the main fluctuations in the data. We improved this explanation on the **3.1 Preprocessing** subsection: *"To provide the CNN model with continuous time series, we performed data imputation using Multiple Linear Regression (MLR). This method is applied only when the wells exhibit similar behavior in their time series, as determined by Euclidean distance. The distance is calculated between GWL time series after standardizing each series to be zero-centered with a standard deviation of 1, followed by detrending to remove linear trends. This approach ensures that the comparison focuses on the primary fluctuations in the data."*

(4) *L 142: Thanks for the clarification. I think it makes sense to add "(15km*15km)" of "pixels (5km*5km each)" to make the scale of meteorological data clear. As information is scattered from above*

We agree with your suggestion. The modification is seen in **Table 1**.

(5) *L230 and Table3: Thank you for addressing this comment. I am still not convinced "high stability" should be equal to "high variance", as for me higher stability means low variability. Thus it is counterintuitive for me. Could you please expand this in Table 3 if appropriate? For example, in the case of the power spectrum the right column provides information on what high versus low values*

> *actually mean. That would be helpful here as well, esp. with this counterintuitive definition. Please also check if this is reflected in the text, as e.g. the correlation coefficient might be interpreted in reverse way.*

We understand your confusion and agree with your intuition. A high variance in the means across segments indeed results in lower stability. Thus, a lower value of this stability feature, as computed by the ***tsfeatures*** package, indicates that the time series is more stable, with its mean remaining relatively consistent across different segments. We understand your concerns and suggest that, for future studies, computing the inverse of this measure would yield more intuitive results. We modified description in **Table 3** to *"Low values indiscate greater stability, meaning that the GWL remains within a certain range without substantial variations or trends"*, and modified the section **Performance assessment** for clarity *"Similarly, higher variance of the means through overlapping windows (as indicated by the stability feature defined by Yang and Hyndman (2020)) may reduce model performance."*