Dear Marvin Höge,

Thanks for your valuable remarks. We carefully reviewed each comment and replied to it accordingly. Firstly, we refer to the main comments on the general evaluation and recommendations, below we give answer to each specific comment.

(1) In the introduction, it is mentioned that some consider machine learning methods as "black boxes". Explainable artificial intelligence (xAI) tools or similarly called methods are supposed to help here. Therefore, a brief coverage of advances in this field – ideally in the field of groundwater research if available - would be worth mentioning.

We acknowledge the importance of addressing the implementation of Explainable AI techniques in our research, and we therefore, integrated the topic in the introduction (line 37) and discussion sections (line 224).

(2) The main point of concern, however, is as follows: Overall, the performance of the employed CNN model, as presented, e.g., in Figure 6, is not fully convincing. Even the well-performing models - according to NSE and R2 - show mainly a sinusoidal pattern with only slsight variations – yet, it is these variations that would be interesting to be modelled. Otherwise, a sine function-based model with a mean trend might often be sufficient and provide the same goodness-of-fit values. Therefore, one can assume that the used model architecture (together with only precipitation and temperature as inputs) is not complex enough to capture more of the dynamics. The subsequent analysis of performance with respect to geospatial and time series features therefore appears to be weaker than it could be. It appears to be difficult to deduce relations between features and model performance if the model does not perform convincingly in the first place. All correlations reported are rather weak with the strongest anti-correlation being -0.62.

Thanks for raising these interesting points. A model that adequately represents the lower and higher periods due to dry or wet years holds higher relevance for groundwater management than catching the small spikes on the observations. We calibrated a sin curve model on all stations for a proper comparison. As a result, a simplified approach such as a sin-function-based model lacks reproducing the dry years with lower GWL and wet years with higher GWL, as shown in Figure 1. Contrarily, the CNN model is able to reproduce the low periods on GWL besides providing a better goodness-of-fit (Figure 2). We selected a different example to show the model fit, since in most occasions, the CNN model does not show a sine pattern and instead try to follow the observed pattern. Additionally, it is important to note that based on the literature, AI models applied to forecast groundwater levels, when using only precipitation and temperature as inputs (especially without including antecedent GWL), are generally lower than models applied for surface water. For instance, Wunsch et al., 2022 obtained similar NSE and R2 values using only precipitation, temperature, and relative humidity as input features. As seen by the review paper Tao et al., 2022, most listed papers make use of the antecedent GWL, which eventually leads to better results. Besides, our model showed robust prediction skills since training and testing differences on mse are not significant (increasing on average 0.02 from training to testing period).
We shortly address the previous discussion topic on the revised manuscript (see line 257).

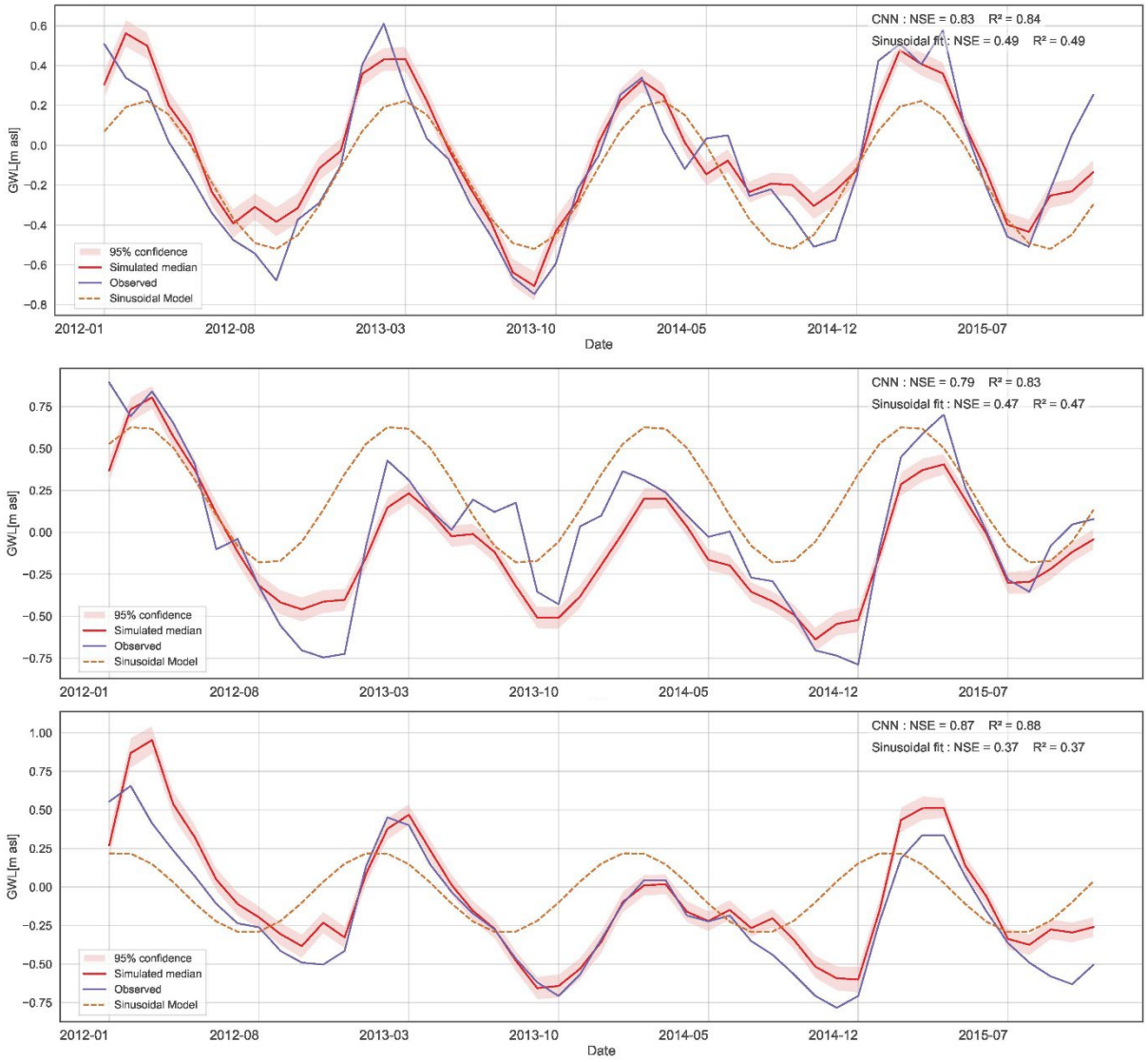

Figure 1. GWL time series observations during the test period, simulated median and sinusoidal model fit.

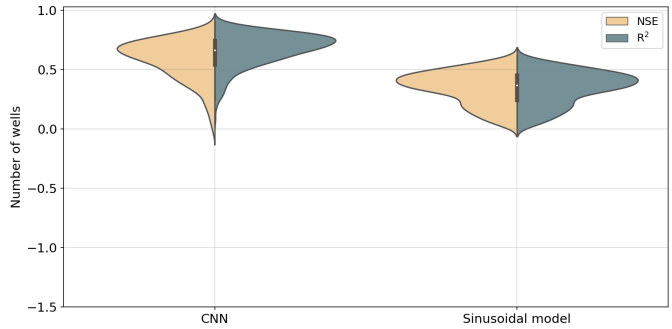

Figure 2. Boxplot of R2 and NSE values of CNN and sinusoidal models.

(3) Along these lines, the used geospatial features are all interesting but the timeseries features are too many and some are hardly tangible. A thorough explanation of their meaning, range of values, etc. would be beneficial. If time series features shall be part of the analysis, I recommend to focus on only a few of the ones provided. Yet, in this case, the question remains: What is exactly gained from relating these time series features to model performance? For instance, a time series can be rated as complex while an adequate model could still be able to predict it.

We agree that some time series features can be removed, so we keep some of them, adding a less technical description and discussing their physical implications when dealing with groundwater level modeling (see revised manuscript table 3, line 165, paragraph starting on line 287). We believe that the time series features can provide insights into the underlying patterns within the groundwater levels. For instance, time series considered complex (e.g., high approximate entropy or number of peaks) contain irregular and unforeseen patterns, increasing the effort of building a generalized deep learning model based only on precipitation and temperature. Time series features might, therefore, justify the need for additional information instead of focusing efforts on implementing multiple DL architectures. We are aware that obtaining additional information such as pumping rates is often challenging, particularly when dealing with a regional scale, as is the case in this study. By extracting patterns on the GWL itself, we can approach ways of understanding the observed patterns even when information is lacking.

Detailed comments and replies:

Note: *please find in black the manuscript comments and in blue the reply.*

Page 1

Line 5: Comment 1: Highlight

Line 5: Comment 2: Highlight

Line 6: Comment 3 Highlight

Line 11: Comment 4 Highlight Agree: changed

Line 12: Comment 5

There are also other measures like NSE and BIAS shown in the results. Here it sounds as correlation was the only performance metric.
Please clarify.
Here, we mentioned Pearson correlations referring to the link between geospatial/time series features not related to the model performance.

Line 12: Comment 6
spelling, see line 10
Agree: modified

Line 13: Comment 7
which mean? Please specify.
Agree: modified: The long-term mean of groundwater level time series

Line 14: Comment 8
Only a conjecture. Not sure this should be stated in the abstract like this.
Agree: modified

Line 14: Comment 9
Please reformulate. Complex is a vague term in this context.
Agree: modified

Page 2

Line 25: Comment 1
This should be a seperate sentence. The "since" implies a reasoning between the two parts that is not clear. Also it should be relocated to suit the red line of the text better.
Agree: modified

Line 26: Comment 2 Cross-Out
Agree: modified

Line 37: Comment 3

To a large part this might be the case, but there are explainable AI tools to help eliciting functional relationships between inputs and outputs. Further of these are develeoped. This should be briefly discussed here or elsewhere in the introduction, because they effectively counteract the "ML = black box" claim.
Agree:added

Line 40: Comment 4 Cross-Out
Agree: modified

Line 45: Comment 5
This is due to the fact that groundwater systems are typically (not e.g. Karst) highly autocorrelated with inertia in their dynamics. Please briefly discuss this in the context of NARX being a suitable tool to capture this and therefore perform well.
Agree: modified

Line 45: Comment 6
which last one? Please clarify whether this refers to one of the shown model types or an implementation in the cited references.
Agree: modified

Line 52: Comment 7
Please rewrite. Suggestion: "it is known that in order to achieve more"
Agree: modified

Line 53: Comment 8
replace by ","
Agree: modified

Page 3

Line 72: Comment 1
I suggest to write Lower Saxony instead of abbreviating it. Using GWL as only abbreviation helps readability.
Agree: modified

Line 79: Comment 2
see above
Agree: modified

Page 4

Fig. 1: Comment 1
please replace by ",000" - K is unusual in the context of maps.
Agree: modified

Line 84: Comment 2
Please add an explanation why this is the case. Probably, among other things, due to more difficult delineations of water protection zones in karstic regions. Hence, there is less interest in pumping there and fewer wells.
Agree: modified

Line 85: Comment 3 Cross-Out
Limits
Agree: modified

Page 5

Fig. 1: Comment 1
Please add figure parts (a) and (b) and clearly mark them in the caption. 1st figure legend: Are there no 1 month gaps in the data ? 2nd figure legend: please use "-" instead of commas for ranges. Why were these ranges chosen like this?

Agree: modified
The ranges are set based on the scheme of quantiles

Fig. 3: Comment 2
This is a great figure! Maybe a bar plot could be added showing the distribution of filter depths of all wells. Yet, this could also go into the appendix but is also no must to provide.
Agree: Figure added in appendix

Fig. 3: Comment 3
please repeat definition of SMI for understanding
Agree:added

Page 6

Line 98: Comment 1
Thanks for highlighting this! Please add a reference to the website were data can be requested/downloaded.
Agree:added

Line 105: Comment 2
Please clarify what this means. Do the used CNN have probabilistic parameters?
Agree:modified
The Bayesian process is used as hyperparameter tuning/ so model optimization (more about this approach can be found in https://github.com/bayesian-optimization/BayesianOptimization).

Line 106: Comment 3 Cross-Out
Agree:acepted

Line 117: Comment 4
This should be split up into with smaller or equal one months gap and one to two months gap. Please see comments for figure 2.
Two months gap means 1 missing value, 3 months then 2 values. We modified this on the text and on figure 2 to avoid confusion

Line 115: Comment 5
unclear, please reformulate
Agree:modified

Page 7

Line 124: Comment 1
D
Agree:acepted

Line 131: Comment 2
Why was data split up like this? Should hyperparamter tuning be based on more data since at the lower end 21 years of monthly data gives 252 data points? 10 % thereof appears too little for hyperparameter tuning.
We intend to provide a significant amount of the dataset for training purposes so the model can learn most patterns encounter in the time series. We also stick to the general practice of using 10-15% of the data for hyperparameter tuning.

Line 131: Comment 3
So all wells provide data until 2021 but start differently? I think this would be a valuable additional information in the data chapter.
Agree:added in the data section. Yes, most time series are available until 2021 but start differently.

Line 136: Comment 4
Please specify, see comments above. There are many Bayesian methods that can be used for optimization.
Agree: added on comments above. Reference added

Line 144: Comment 5
Is it significant or minor degree? Generally, this sounds like a result and should not be stated as such here. Please remove or adopt.
Agree: modified

Page 8

Line 148: Comment 1
Unclear, please rephrase
Agree: modified

Line 152: Comment 2
Please specify: is this a 1km radius around the well? Or is this a grid cell around the well. Buffer as term is not specific enough.
Agree: modified. Refers the 1km radius around the well.

Line #: Comment 3
Please specify: this referd to a long internal list of the used packages and not to the - not so long - list of features in Table 2 so Table 3 should be mentioned earlier to avoid confusion. How do you assess the relation between correlation between "metrics and added value to the analysis"?
Agree: clarification added, it refers to a long internal list on the Python packages. We saw that the correlation is significant and that the feature could potentially be linked to the physical understanding of GWL levels.

Line 154: Comment 4
This seems like too many metrics and for all those that are used, an explanation has to be added to provide context for the reader of how to interpret them. The descriptions in Table 3 are not enough for this since they only provide a brief intro into what each metric resembles but not about what a resulting number could mean.
Agree: modified.

Line 158: Comment 5 Cross-Out
closer
Agree: modified.

Line 159: Comment 6 Cross-Out
observed data
Agree: modified.

Line 161: Comment 7
Please elaborate why? It is not clear why a measurement for the time series complexity or flat spots should have any relation to model performance. Sure, a more complex timeseries might be harder to match with a model hence lower NSE or R2 might be expectable. But which additional value is gained from getting the correlation values there?
Refer to answer on comment 3.

Table 2: Comment 8
Check consistency as for other places in the text: 1 word or 2 words
Agree: modified.

Table 2: Comment 9
what about distance to next water body (lake or river)?
At first, we added this feature, but we did not found a statistically significant correlation with model performance. Still, we believe this variable might be relevant but only for a subset of the wells in the study area.

Page 9

Line 166: Comment 1 Cross-Out
fit
Agree: modified.

Line 173: Comment 2
I overall agree on the seasonal pattern - but the up and down throughout the year is not the difficult part in GWL predictions. It is what this up and down looks like exactly. And this is hardly met, even within the confidence intervals. This is indicated in the following sentences but should be directly tied to Figure 6 a or b - especially w.r.t. b: Meeting a

pattern occasionally or by chance seems to be no solid foundation for a subsequent analysis between time series features etc.
Refer to answer on comment 2.

Line 173: Comment 3 Highlight

Line 174: Comment 4 Cross-Out
Agree: accepted.

Page 10

Page 11
Line 178: Comment 1
The plot including the maps is nice. Yet, please increase font size of legend entries.
Agree: modified

Page 12

Line 181: Comment 1
Here, these are relatively high correlations, but generally, 0.4 and 0.3 are not considered high correlations.
They might be low correlation when dealing with surface water data but in groundwater time series where the influential variables are less seen, these correlations (which are still statistically significant) get more value.

Line 184: Comment 2
lower metric values or correlation between them and the features? Please elaborate what is meant here exactly.
Agree: modified.

Line 186: Comment 3 correlate with? (see comment above)
Agree: modified.

Line 188: Comment 4
Isn't this counterintuitive?
A time series with high autocorrelation can lead the model to a poor generalization, making it unable to learn patterns through the CNN filters. Consequently, low performances can be observed during the testing period.

Line 189: Comment 5
conjecture? Then please explain why?
Because if GWL on the previous step is included, then it would be one of the most influential input features on the model output (See Figure 2 of Chakraborty et al., 2021, and Zhang et al., 2023). If additionally the time series is highly autocorrelated, then the current state can be better explained by the previous one.

Line 193: Comment 6
why should there be a correlation between goodness of fit and time series length at all - except for the amount of available training data?
It is true that still there is not a clear correspondence between amount of data and model performance. However, the more data is available for training, the more patterns and relationships can be learned by the model. To avoid overfitting, more validation data also should be available.

Page 13

Line 199: Comment 1
Please elaborate what this means here. This sentence part seems out of place.
Agree: modified.

Line 210: Comment 2
precision of the models is generally high as can be seen with the narrow uncertainty bounds in figure 6. The accuracy seems to be the issue in several cases. Please reformulate and specify.
Agree: modified.

Line 221: Comment 3

As written above, please specify which Bayesian procedure was applied - a Genetic Algorithm can be Bayesian as well. Therefore this generalization cannot be made.
Agree: modified.

Line 256: Comment 4
Yet, how shall clear relations to features be derived it remains uncertain whether the used model actually captures as much information as it should to allow such a subsequent analysis?
Linked to main comment (2)

Page 14

Line 229: Comment 1
This should be shown somewhere. So far it is just claimed.
This is discussed on the performance evaluation. Here is stated as an introduction to the further subsections.

Line 238: Comment 2
Why would it hinder the relation? The deeper aquifers are simply not part of this study. This is fine, but the reasoning here seems unclear.
It is true that regionally deep aquifers are not consider in this study. However, here we refer to the relative depth as seen in Figure A.1.

Line 250: Comment 3
Yes, but a deep learning model could learn this since it also has a seasonal occurance throughout the vegetation period. I think this rather points towards that the model should be more "complex" in order to capture such dynamics. And this is not sufficiently discussed so far.
The DL model based on P and T as inputs can capture many interactions but not entirely, even with a more complex model. The interception process involves many other in-situ characteristics (soil type, land cover, human-induced activities) that do not feed the model in this case. Introducing input variables such as evapotranspiration is seen as a more appropriate approach.

Line 258: Comment 4
Please specify, this is too vague.
Agree: modified.

Page 15

Line #: Comment 1
evidence is a strong word here. With the low correlation values given, it "indicates" at most. Even then, with all the other confounding variables discussed before, I think this conclusion is not convincing.
Agree: modified.

Line #: Comment 2
Again, why should there be one?
See comment above. The more data is available for training, the better the DL model can learn the patterns and relationships between the input-output. More data on for validation or tuning also help to avoid overfitting.

Line 275: Comment 3 Cross-Out
are
Agree: modified.

Line 279: Comment 4
The analysis to geospatial features is interesting and provides a foundation for further investigation. The analysis of the intrinsic time series features is not as convincing yet - to a large part it remains questionable why a certain value of such a feature should give us insight into whether a model performs well or bad. Like the rest of the paper, these conclusions should therefore rather focus on the geospatial features and expand this to draw more reliable conclusions.
See answer to main comment 3

Page 16

Line 293: Comment 1
CNN in uppercase?

Agree: modified

Page 17

Line 326: Comment 1
Link does not work
Agree: modified. Link should work now.

Dear Jonathan Frame,

We appreciate your valuable inputs regarding the manuscript and carefully replied to your comments.

(1) This paper would greatly benefit from additional description of the training procedure and evaluation when dealing with gap-filled or processed data. Another benefit would be further evaluation of the sensitivity of model performance to the gap filling and data processing measures."

Additional details of the gap-filled process were added to the revised manuscript (see section 3.1). Regarding the training procedure and sensitivity of model performance to the gap filling, we believe that it can be helpful in further and future research. Still, by only using time series with good data quality in terms of data gap lengths and frequency, we seek to avoid major influence of data imputation approaches. Therefore, we think that a sensitivity analysis is a bit out of the scope of this study and could be focussed on in a follow-up analysis.

(2) Line 44: These claim should be cited: "In terms of accuracy and calculation speed, the CNN models outperform the LSTM. NARX models performed, on average, better than CNN."

Agree: modified (revised manuscript line 46)

(3) Line 47: "Most studies have successfully applied these techniques for GWL forecasting using only meteorological variables as inputs." You might be interested in this paper: Gholizadeh et al., "Long short-term memory models to quantify long-term evolution of streamflow discharge and groundwater depth in Alabama" Science of The Total Environment Volume 901, 25 November 2023, 165884, where the did in fact include site geospatial characteristics to make predictions of wells that were held out from the training set (ungauged).

Agree: we included the reference in the introduction (see revised manuscript line 57).

(4) Line 116: Can you please provide the total number (and percentage) of gap filled values referred to here: "To provide the CNN model with continuous time series, we performed a data imputation process through a Multiple Linear Regression (if enough dynamically similar wells based on the Euclidean Distance". Can you also explain if these values were removed, or should be removed, from the loss during training, and also removed from the evaluation?

Agree: We included precise numbers regarding the data gaps (see revised manuscript line 124). We did not remove them from the training phase since the number of filled values is not considerably high.

(5) Line 120: Similarly can you provide the total number (and percent) of data points that were modified as outliers described here: "We removed these anomalies by

finding the highest slope in the cumulative sum"? Is this a standard approach? I don't think this description is satisfactory. I see from your code that you identify these based on "initial point where the values increase by 0.5 of the standard deviation". This is an important point that should be explained in the paper, as well as the decision to use this processing method.

Agree: We added further explanations to the revised text (see line 128).
Only 28 wells were identified to have jumps/steps on the temporal record. The cumulative sum is commonly used to detect changes in the mean or variance along the time series and is not referred to as outliers. Here, we intended to detect jumps/steps on the observed values that can hinder the model training or might introduce confusion to the model due to potential changes in the dynamic of the groundwater levels. The optimal fraction of standard deviation was determined through trial and error by visually inspecting the detections and selecting the value that best adjusted to most jumps.

(6) **Line 135: Is it really necessary to give the equations for r-squared and NSE, as you don't provide the equations for MSE or BIAS? There is also more unfamiliar calculations made in Tables 2 and 3 with no equations provided, and also the main CNN model is not described with equations. I guess I would suggest just removing equations 1 and 2, avoiding an asymmetry in descriptions, rather than adding equations for all the rest of the calculations.**

Agree: equations were removed.

(7) **Line 176: "Occasionally, in poorly performing models, the pattern of the GWL observations has been generally learned but with a strong Bias." This is a little concerning, and I think it would be work describing in more detail. Similar to your NSE/r-squared cutoffs above, can you provide a quantification of these problematic BIAS wells, something like in how many wells does the prediction not intersect the observation? What causes this BIAS, is it an unusually high section in the training period? I wonder if there is anything in the data preprocessing that plays into this issue**

Around 10% of the wells show strong bias (>0.3), meaning the model has little or no intersections with observations. Differences in spatial resolution between the input data (gridded precipitation and temperature) and the GWL observations can cause this bias at some stations. The data processing results in less than 5% filled values on non-complete series; therefore, it is not likely this is the cause of high bias. This is added in the discussion (see line 253).