Dear Marvin Höge,

Thanks for your comments on the current state of the manuscript. We carefully revised them and carried out the necessary modifications.

*Overall, the manuscript shows improvements, most comments have been addressed (e.g. refined explanations and the added figure to the appendix) – yet, when it comes to the major points of concern not always fully satisfactorily (see points below). I still think this paper can be a valuable contribution to the field of water research, but at the same time it should also not oversell outcomes – in particular w.r.t. to the analysis of correlation coefficients. That said, it is no flaw of a study to say that a particular method was applied but did not reveal clear relations. However, it has to be clearly stated. Then, the community learns more about the capability of the employed methods. I am open to recommend publication but not in its current form. I suggest another round of iteration to address the following points:*

*1) Most importantly, the results section needs to be partially rewritten:*

*l. 198-199: "Although correlation coefficients are statistically significant."*

*-> Were significance tests conducted to support this? If so please show the results (e.g. in an appendix section). Otherwise, please rephrase.*

We report only those correlations that demonstrate statistical significance, ensuring they fall within a 90% confidence interval (p-value <1) to ensure the reliability of our findings. Please refer to lines 197-198 in the re-revised version of the manuscript

*l. 200ff: "The R2 increases as the distance to the coastline does, whereas the bias reduces."*

*-> There is no clear trend visible*

We agree there is not a clear trend visible, even though Pearson correlation suggests a positive relation between model performance and distance to the coastline. Therefore, we modified in text to: "Although there is no clear spatial pattern followed by R2 and NSE, the Pearson correlation suggests that model performance improves with increasing distance from the coastline." (see lines 220-221 in the re-revised version of the manuscript)

*"The proportion of the most common land cover type in the study area (non-irrigated arable land) relates positively to model performance. Conversely, wells with a significant surrounding area of forest or high LAI display lower correlations. Sink and low relief…*

*-> The correlation coefficients to support these statements are in the area of +-0.1x. This is not sufficient to support these claims.*

Thanks for your comment. We modified the text to account for the limited certainty of this interpretation: "The proportion of the most common land cover type in the study area (non-irrigated arable land) suggests a positive relationship with model performance. Conversely, wells surrounded by significant areas of forest or high LAI tend to show lower correlations. Sink and low relief areas with medium to high SMI may negatively affect performance. Hilly regions might indicate lower accuracy, while areas with high drainage density or a high topographic wetness index suggest better model performance." (see lines 221-225 in the re-revised version of the manuscript)

*L 205ff: "Stronger correlations, mainly negative, are found for the time series features. Overall, we found that autocorrelation reduces model performance."*

*-> Also the results part of the time series features analysis has to be rephrased. First, the correlations might be stronger, but starting from essentially no correlation and reaching up to 0.52 maximally is no strong correlation. Therefore, formulations should be appropriate and reflect that relations are indicated but generally not clearly supported.*

*Overall, regarding the correlation coefficients it should be clearly stated that most features showed no clear correlation.*

*Reporting such results properly is no bad because it shows other researchers what one can and cannot expect from the applied CNN + feature analysis procedure used in this paper. The reduction of considered features in Fig. 7 and the corresponding text is already an improvement. Yet, it has to be made clear that the conducted correlation analysis did mostly show no clear correlation. This does not undermine the content of the paper but is proper scientific conduct.*

We agree with your suggestion. Modified in text to "Regarding time series features, autocorrelation may reduce model performance. This might not be the case when using antecedent GWL as an additional input feature, where GWL shows the highest influence on model output (Chakraborty et al 2021), better explaining the current state based on the past one if the time series is highly autocorrelated. Similarly, higher time series stability (higher variance of the means through overlapping windows) may reduce model performance. Increasing flat spots and long strikes above or below the mean are negatively correlated, particularly with the NSE metric. Positive correlations are mainly associated with complexity measures such as approximate entropy and the number of peaks. The time series length positively correlates with R2 but does not correlate with NSE. Higher values of the Fourier power spectral density at one year (indicating stronger annual seasonality in the observed GWL) result in higher model performance." (see lines 227-234 in the re-revised version of the manuscript)

*2) Fig. 6: Thanks for comparing the CNN results to a sinusoidal fit. Yet, first, I was referring to a "a sine function-based model with a mean trend" as a (too) simple reference model. A simple sine function alone clearly cannot follow the observed pattern of GWL. The model I was referring to would look something like this: y(t) = m(t) + a*sin(bt+c) with time t and parameters a, b, c of the sinusoidal part. The "trend" part m(t) could be something like a moving average or regression model to describe a trend that depends on the precipitation of the last 9 months or so. With such an approach, I think this would really have led to a simple reference model that could also be reported in the appendix. The sole sinusoidal fit obviously cannot be a proper alternative.*

Thanks for your inputs. A sin model defined by y(t) = m(t) + a*sin(bt+c) with time t and parameters a, b, c of the sinusoidal part and m(t) defined as the precipitation trend of the last 9 months (linear regression) is now included. Figure 6 shows the original examples with both CNN and the new baseline model.

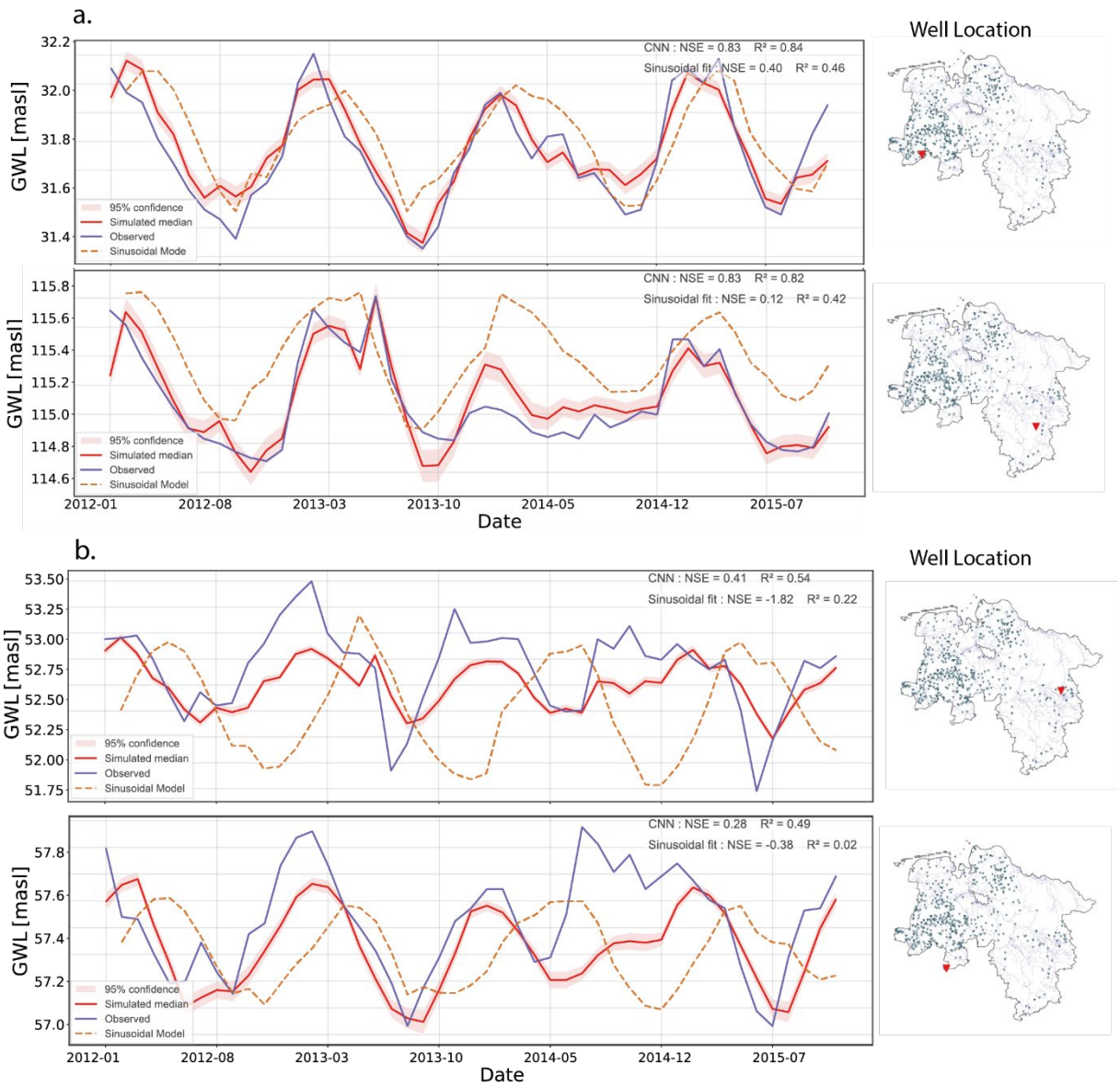


*Figure 1. Examples of the observations, CNN model, and baseline model (sinusoidal plus precipitation trend) for cases of (a) high performance and (b) low performance.*

*Second, results for a reference model should have been reported for the same GW stations as shown in Fig. 6. It seems like only the (well-fitting) first panel of Fig. 6a was reported in the author's reply while the other two panels in the reply do not match the original time series in Fig.6. I leave it to the authors to decide whether my original suggestion should be pursued.*

Yes, they are both different figures since we tried to select the most representative but still within the highest and lowest performances. We kept the same time series examples on this version of the manuscript.

*3) L 224-227: "To better interpret the non-linear behaviour between groundwater and its 225 influencing factors, some studies applied explainable AI techniques (Chakraborty et al., 2021; Zhang et al., 2023). However, this implies including all the analyzed features as inputs on the model, most of*

*which correspond to static data (as regionally accessible information) that might not add value to the sequential model."*

*-> These are newly added sentences. The second sentence is not clear and sounds like a justification of why xAI methods are not applied after drawing attention towards them in the first sentence. I think it is fine that this topic is beyond the scope of this paper but I suggest not to add it to the Discussion section then. This could go into the Conclusions/Outlook section, saying that such techniques might be able to shed more light on relations that could not be identified via the employed methods here.*

Thanks for your comment. We removed from the mentioned paragraph and added to the conclusions: "Nonetheless, incorporating explainable AI techniques in future studies is recommended to enhance the interpretation of the non-linear behaviour between groundwater and its influencing factors." (lines 243-244 in the re-revised manuscript).

Dear **Referee #3,**

We appreciate your careful and detailed feedback on the current study. We understand your new and fresh view on the manuscript after the first round of revisions. Your suggestions were revised in detail as they bring up additional modifications.

**General comments**

*(1) The abstract does not really flow well, some terms and sentences are not clear, please revise to improve clarity. The temporal data resolution should be added as well. The clarity is also an issue in the main text, I indicated some examples, but please generally check the clarity as this is really important for the readers*

Thank you for commenting on the abstract. We agree that it should be improved for clarity, and we have added the temporal resolution. (See lines 5-17 in the re-revised version of the manuscript)

*(2) This is also the case for the title, which does not really match the manuscript. From my reading, you did not investigate performance of the features but linked performances to the features which is different in my view. Please revise*

We agree with your comment on the title and therefore modified to "Assessment of Groundwater Level Modelling using a 1D-CNN: Linking Model Performances to Geospatial and Time Series Features".

*(3) I am not sure why the authors train one model per site instead of one model for all sites which could perform much more robustly and avoid overfitting? Also seeing in the context of the current comment by Kratzert et al. on improved machine learning models when training on multiple catchments, could you comment on that, please? The motivation for such an approach should be presented to the reader.*

Thank you for bringing this up. We are aware of the paper by Kratzert et al. (2024) and acknowledge its conclusions. However, we recognize that findings in river run-off modeling are not always easily transferable to subsurface groundwater contexts. We are also aware of the recent publication by Heudorfer et al. (2024) about global groundwater level models which shows results where global models can be indeed better than single trained models but the difference is by far not as big as in the paper by Kratzert et al (2024) .Up to now, the single well per model setup is the more widely used one and to be considered the state-of-the-art in groundwater applications, which makes our approach more relatable to these setups. We added this clarification to the re-revised manuscript on lines 48-51.

*(4) The authors miss to reflect their model performances in context of overfitting a typical issue that can occur in highly parameterized machine learning applications. The authors did not present performance from training and validation and reflect on potential mismatches in performance compared to test data. Overfitting can also cause low model performance on the test data, however, the authors only discuss poor model performance in relation to physical and time series properties. Also, I am missing information on the covered time periods per stations in addition to the time series lengths to better understand the variable coverage of training, validation and testing periods. The reader is left with incomplete information to fully interpret and use the presented results.*

We agree that overfitting can affect model performance on the test data and should be included in the discussion. Therefore, we have added Figure 1 to the re-revised version of the manuscript appendix (see Figure A3), which displays the difference in model performance (RMSE) between the validation and testing periods. This figure highlights the differences, which are below 0.20 in most cases, as an indicator of overfitting. Most sites with the highest differences also show low model performance, as seen in manuscript Figure 5.

Respective elaborations can be found in lines 213 in the re-revised manuscript

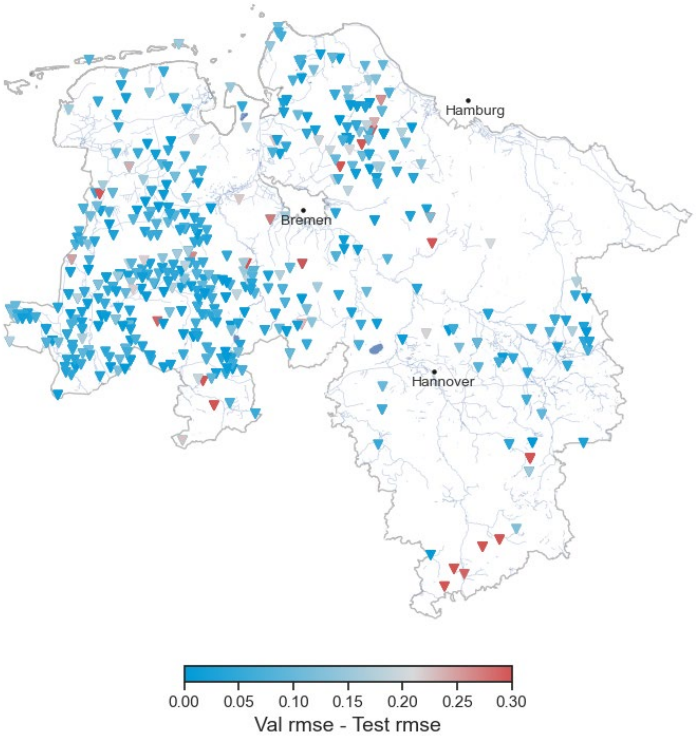


*Figure 2. Difference in model performance (RMSE) between validation and testing periods*

To provide additional information to the reader, we included the following plot on the appendix. It displays the time range of the GWL time series. (Figure A1, referenced in line 94)

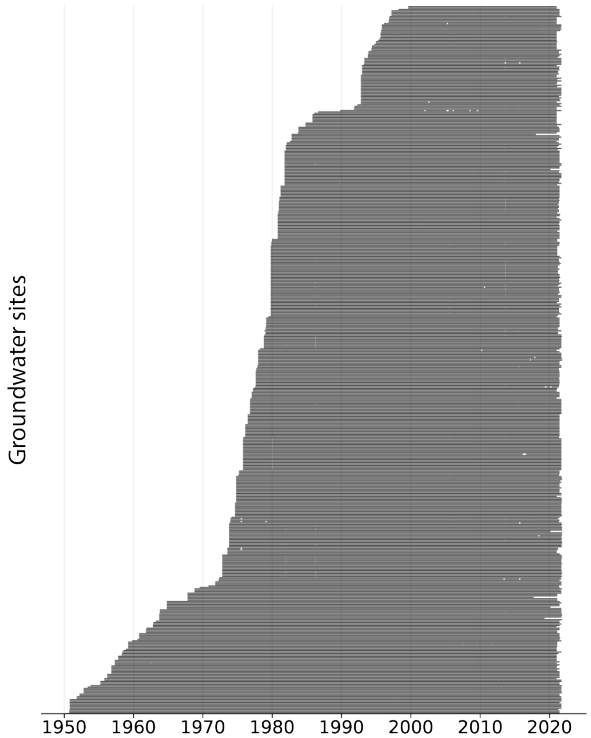


*Figure 3. Time range of GWL observations. The blank spaces correspond to missing data.*

*(5) I agree with a previous comment by Marvin Höge who brought up the point of interpreting model performance and suggested to compare CNN performance with a simple model (.e.g sinusoidal) to see*

*how much a complex CNN actually adds to the predictability. I did not see that the authors really addressed this issue, and the response partly missed this issue.*

Thanks for commenting on this. We added the base-line model suggested by Marvin Höge for comparison. Please refer to comments 2 in our response to Marvin's review.

*(6) In relation also to the above issue: How much does your result depend on the selection of "relevant" features? In the discussion you wrote that it is hard to separate the effects, but actually maybe you would also miss effects. Does it also depend on the data quality, length of training, validation, testing periods? What about trends?*

We selected geospatial features based on data availability and their potential impact on groundwater level observations. However, as highlighted by Tarasova et al. (2024), the lack of agreement on evaluating hydrological catchment descriptors hinders consensus on what is considered as relevant features, in particular for subsurface characterization. Despite this, we aimed to include as many features as possible. Nevertheless, some features did not significantly correlate with model performance and were consequently excluded. (See lines 61-63 in the re-revised version of the manuscript)

The results are indeed influenced by data quality, which is why preprocessing was a crucial first step. The impact of the length of training, validation, and testing periods on the results is consider low. As shown in Figure A4 (density function), reducing the testing period to 3 years or increasing it to 5 years does not lead to major changes in terms of model overfitting. Exploring these aspects in greater depth is beyond the scope of the current study. However, we acknowledge that trends in data and other factors like temporal resolution might also play a role.

We added these clarifications to the re-revised manuscript at lines 213-214

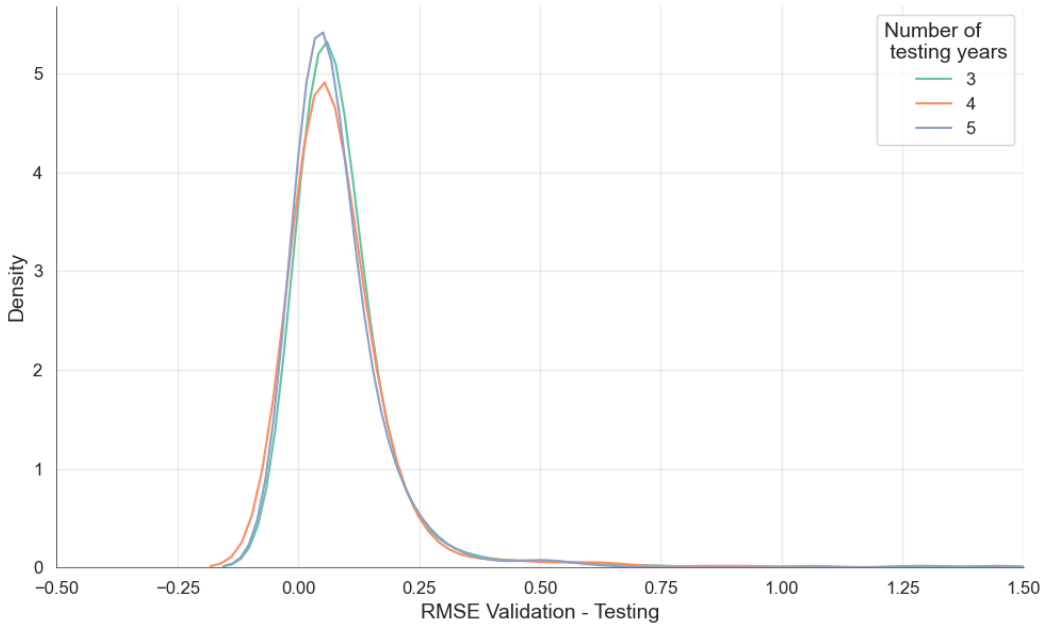


*Figure 4. Model performance (RMSE) difference between validation and testing period for 3, 4 and 5 testing ranges.*

*(7) I assume pearson correlation might not be appropriate, I would not expect a linear relationships between strongly skewed and partly intercorrelated variables. Maybe, it would be also good to provide an overview (and the data, see next comment) of represented feature values.*

We also used Spearman rank correlation but did not observe higher correlations, so we decided to keep using Pearson correlation (we also added a note on this to the re-revised manuscript in lines 181). Additionally, we inspected the relationships visually, which you can see in our GitHub repository (Jupyter Notebook: 04. Geospatial_Accuracy_plot.ipynb).

(8) Please add more complete information on the data used in this study, e.g. provide links to the download pages, potential download criteria, download dates, versions etc. The code repository is great, however, it is still not fully reproducible without the input files unfortunately. At best, could you also include the data into the code repository. This would strongly increase reproducibility of your results, potential follow-up studies and be highly appreciated. At least, the processed geospatial and time series features data and metadata of the included stations should be possible to include.

Thanks for your suggestions. We have added the groundwater level dataset (raw dataset, gap-filled dataset and corresponding shapefiles) and meteorological forcing (HYRAS data corresponding to each well) used as input for the CNN model to Zenodo in reference (Gomez, 2024) in the re-revised manuscript. Additionally, the sources for the geospatial information are noted in Table 2, the link to each source it is joint to the citation. We hope this enhances the clarity and usability of our data.

**Abstract**

L5 This is not clear “Likely causalities of this discrepancy”

We agree it is not clear and have modified the text accordingly. “The reasons behind this discrepancy in model performance have been scarcely examined in previous studies”. (see lines 4-5 in the re-revised version of the manuscript)

L6 how do you quantify the “effects of …” This term does not match with what was done

We agree with your point and consider 'effects' is not the appropriate word in this case. We have modified it in the new version to: “Here, we explore the relationship between model performance and the geospatial and time series features of the sites.” (see lines 5-6 in the re-revised version of the manuscript)

L11 why do you use pearson correlation, are you assuming a linear relationship? I think this is very unlikely actually, have you checked the relationships visually also?

Thanks for your comment, please refer to respond of main comments (7)

L14 “exhibit better metrics” this is confusing as metrics could be anything and you are probably referring to model performance, so I suggest to also write that.

We apologize for the confusion and modified in the new version to: “Besides, GWL time series containing more irregular patterns and with a higher number of peaks might lead to higher model performances, possibly 15 due to a closer link with precipitation dynamics”. (see lines 13-15 in the re-revised version of the manuscript)

L16 “external physical factors” not matching time series properties

We agree and modified in text to: "…our work provides new insights into how geospatial and time series features link to the input-output relationship of a GWL forecasting model.". (see lines 16-17 in the re-revised version of the manuscript)

**Introduction**

L20ff I think the authors should try to cite peer-reviewed publications instead of citing an UNESCO report (which is also not fully referenced in the literature) for the first 3 sentences. Moreover, I find it a bit irritating that the authors start their statement with water use, although they want to model groundwater levels, I think this paragraph could be largely improved.

Thanks for raising this concern. We modified the first paragraph on the introduction section and removed UNESCO citations (see lines 20-28 in the re-revised version of the manuscript)

L25 This seems incorrect "approaches based on groundwater observation sites"

Thanks for your input. We modified the first paragraph of the introduction section. (see lines 20-28 in the re-revised version of the manuscript)

L49 Is it necessary to cite the preprint? I think it is the same publication as Wunsch et al 2021. Also for Wunsch et al. 2022 the link leads to the preprint not the actual published paper, please revise your references in that sense.

We apologize for the confusion and modified the reference on the new version. (see lines 57, 540, 543 in the re-revised version of the manuscript)

L65 what are "relevant" features? Please, revise.

Thanks for your comment, please refer to respond of main comments (6)

L84 "through"? Please provide the exact information to access the data, at the best an open data portal where the analyzed data is easily accessible.

Thanks for your comment, please refer to respond of main comments (8)

L85 Is the data directly measured monthly or are these already aggregated values over smaller time intervals?

Thanks for your questions. No aggregation was needed since we received the data on monthly resolution as specified in line 99 of re-revised version of the manuscript

L95 what is the "soil moisture index (SMI)"?

The Soil Moisture Index (SMI) is a measure developed by Hunt et al. (2009) to determine how wet or dry the soil is at any given time based on the minimum and maximum moisture levels that the soil can hold. We added an explanation regarding SMI in the re-revised version lines 101-103. Additionally, a description for geomorphological categories are in the source legend.

L102ff previously you mentioned that meteorological variables were provided by the State authority. Please, check

We apologize for the confusion. The meteorological variables (precipitation and temperature) are extracted from the German Weather Service (DWD). This dataset (HYRAS) is based on the publications Rauthe et al. (2013) and Frick et al. (2014)

L111 “resampling” is unclear what is means exactly. Please, be specify.

We agree it is unclear. In this case, it means resampled from daily to monthly resolution. We modified in text for clarity to: “Simultaneously, the meteorological variables are extracted per well location and resampled from daily to monthly resolution.” (see lines 117-118 in the re-revised version of the manuscript)

L117 It would be good to provide the versions used, especially for tensorflow

We agree on including the versions to improve reproducibility. We modified the text to: “To achieve the objectives, several Python libraries are used: Pandas 2.0 (Reback, 2020), Numpy 1.23 (Van Der Walt et al., 2011), Scipy (Virtanen et al., 2020), Matplotlib (Hunter, 2007), Geopandas 0.14 (Jordahl et al., 2020), and Tensorflow 2.7 (Abadi et al., 2015) as the most relevant throughout the process.” (see lines 122-125 in the re-revised version of the manuscript)

L122 “exclude wells under strong anthropogenic influences such as pumping” How did you do that? This requires an explanation. This is also highly relevant for your later interpretations of anthropogenic impacts.

We apologize for any confusion and agree on the relevance of this pre-selection for later interpretations. This pre-selection is based on the classification performed by Wriedt et al. (2020), and we received the dataset already classified accordingly. The methodology relies on the agreement between theoretical and observed hydrographs, as well as visual indications of anthropogenic influences. We have enhanced our explanation concerning the pre-selection process. “The initially available GWL information consists of 962 wells. A pre-selection was done based on the categorization performed by Wriedt et al. (2020), which considered the agreement between theoretical and observed hydrographs, as well as visual indications of anthropogenic influences. This process aimed to exclude wells under strong anthropogenic influences, such as pumping, to better capture the dependency between meteorological input features and observed groundwater levels.” (see lines 127-130 and xx-xx(discussion) in the re-revised version of the manuscript)

L125ff this is not clear, please add what you consider as similar and how MLR was done. Would also good to provide references to methods that you used, e.g. PCHIP. Why piecewise? What does it mean, why is that not used for all gaps? There is also a typo “Otherwise”

Thanks for commenting on this. Similarity is measured using Euclidean distance. To ensure that wells with similar dynamics are used for data imputation, a Euclidean distance matrix is computed, and the wells with the closest distances (below 10 percentile of the smallest distances) are selected. Multiple Linear Regression (MLR) is then implemented with these selected wells, ensuring a model R2 score above 0.7.

Although the Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) is a robust method, we opted to avoid it for the imputation of larger gaps. PCHIP is a simple and efficient approach that has advantages over other methods, such as linear interpolation, by preserving monotonicity and natural trends in the data, avoiding unrealistic oscillations, and providing a smoother, continuous first derivative (Fritsch & Butland, 1984).

Reference and modifications are added on the re-revised manuscript at 129-143.

L133 “3 x 3 pixels“ does that mean you use meteorological input of 15 by 15 km given a resolution of 5km? Isn’t that quite large?

We agree that it is a large scale. However, we followed the best practices recommended by DWD.

L135ff Please provide more information instead of mainly referring to Wunsch et al. 2022. As this is your main methodology, these steps should be clear. Also I have not seen any introduction of a 1D-CNN

Thanks for your suggestion. We modified the text to include additional details about the methodology. (see lines 145-153 in the re-revised version of the manuscript).

L161 Why 1 km radius? What are relevant categories?

Thanks for commenting on this. We recognize that the radius of influence can vary significantly between aquifers. However, as demonstrated by Knoll et al. (2019), a 1 km circular buffer can effectively describe the contributing area of a monitoring site, especially when detailed information about groundwater conditions is lacking. We modified the text to: "Regarding categorical variables, the proportion of a 1 km radius around the well is taken as it has been shown to adequately represent the contributing area of a monitoring site, especially when detailed information about groundwater conditions is lacking (Knoll et al., 2019)." (see lines 175-177 in the re-revised version of the manuscript).

L164 which metrics? How do you evaluate the added value?

Here, we refer to the model performance metrics (R2 and NSE). We clarify this in the text as: "A selection is made from the long list of features (available in each package) according to their Pearson correlation coefficient in relation to the model performance metrics (R2 and NSE) and the added value to the analysis (interpretability in the context of groundwater level)" (see lines 178-180 in the re-revised version of the manuscript). Please refer to response on comment (6) for more detail.

L 151 what are sub-sequences?

We referred to the window represented by the sequence length. (see lines 147-149 in the re-revised version of the manuscript)

L187 "where the density of wells is higher." Does that matter if you are training one independent model per well?

We apologize for the confusion. The sentence was removed in the re-revised version of the manuscript.

L189 I cannot see or confirm that, I would suggest to add a scatterplot to observed versus simulated across models to supplements

We apologize for the confusion. As you suggested, we added a scatterplot of simulated vs observed GWL for the 505 sites (Figure 4) to support our statement. We decided to modify in text to: *"*After visually comparing most of the CNN models with GWL observations, a degree of agreement can be noted between the simulated and observed GWL*"*. See Appendix Figure A5. (see lines 207-208 in the re-revised version of the manuscript)

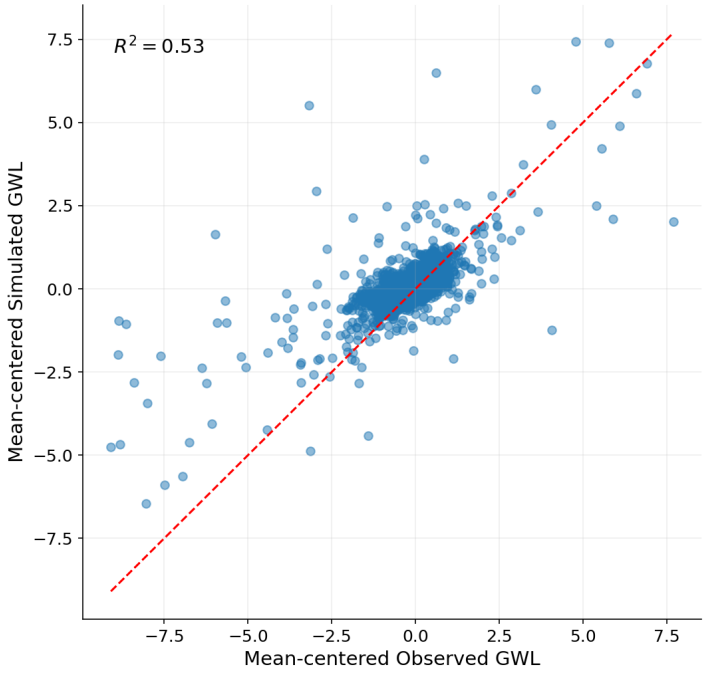


*Figure 5. Scatterplot of simulated vs observed values for the 505 wells for the test period.*

L193 “local variations from the main seasonal behavior are ignored.“ not clear to me what this means

We refer to the local peaks or dynamics that the GWL time series displays at smaller time ranges (e.g. 2 months). We simplified in text for clarity to: “In most cases, local variations on the time series are ignored.” (see line 210 in the re-revised version of the manuscript)

L209 “higher mean variance“? This was not how stability was defined, I think

Stability in this context and as defined in Table 3 refers to variance of the means through overlapping windows of 10 values. We modified the text for clarity to “Similarly, higher time series stability (higher variance of the means through overlapping windows) may reduce model performance.” (see lines 229-230 in the re-revised version of the manuscript)

L229 “are better observed at a weekly or even daily temporal resolution instead of the monthly time step” so what does that mean for your study? It is not clear what you want to say with this information. Actually, the paragraph started with uncertainties, but the arguments presented are not linked back to this issue.

We aimed to highlight that aquifers with unique characteristics, such as karst aquifers or aquifers with strong secondary porosity, require finer temporal resolution to accurately capture their dynamic patterns. This need for higher resolution introduces additional uncertainty when using a monthly time step, as our study primarily does.

We modified in text for clarity to “Certain patterns in groundwater dynamics, especially in karst aquifers or those with strong secondary porosity, become more evident at weekly or daily time steps. Consequently, the use of a monthly resolution in our study may not fully capture these dynamics.” (see lines 245-247 in the re-revised version of the manuscript)

L250 I am not sure what you mean here “understanding the influence of geospatial and temporal features related to the GWL” This seems grammatically incomplete. Please revise. Also, for curiosity, wouldn’t the best model also be best suited for “understanding the influences of … “, i.e. the network structure with optimally tuned hyperparameters?

We agree and completed the sentence to: “However, in this study, our priority is to understand the link between GWL and geospatial and time series features rather than focusing on optimizing the network architecture”. (see clarification in lines 268-269 in the re-revised version of the manuscript)

L260 I do not understand “Every model that could not correctly learn from meteorological inputs might be treated independently.” What is “correctly”, what do you mean with “treat independently”.

Thanks for mentioning this. By "correctly," we meant achieving accurate predictions and “treat independently" refer to analyzing these models separately to understand external drivers (such as local abstraction rates) that might cause their poor performance. We modified in text for clarity to : “Models that do not accurately learn from meteorological inputs might be treated independently.”. (see lines 278-279 in the re-revised version of the manuscript)

L276 I think such an interpretation is critical if most of the area in non-irrigated – how many wells are within irrigated land actually? See my major comment

The current study does not include any wells on irrigated land. Figure 3 displays all the categories corresponding to the analyzed sites.

L290 “variability in climate“ fits better in my view or do you refer to climate change? Then I would not agree though.

We meant variability in climate and modified in text to “This might indicate an aquifer that is less responsive to climate variability, which is often the case with large aquifers”. (see lines 310-311 in the re-revised version of the manuscript)

L293 “trends“ I would be careful with this term - again variability fits better. I don’t think you could extrapolate into trends as model was not trained for this and it is also not the issue discussed here.

Thanks for bringing this up, we agree and use another expression in this case. We modified in text to : “Thus, even if there are upward or downward changes in precipitation, the observations of GWL do not exhibit similar patterns.”. (see lines 312-313 in the re-revised version of the manuscript)

L300 Would a highly seasonal time series be considered as complex? Maybe the term complex is confusing, as I would envision a more erratic and highly irregular pattern, not a regular variability.

You are correct. A highly seasonal time series with regular variability and without significant inter-variations would not be considered complex. However, precipitation time series often exhibit irregular and erratic patterns. Consequently, when there is a strong link between precipitation and GWL, the later will also reflect these complex patterns. This is what we intended to highlight in the text.

L376 “Fernando Nogueira“? Please check

Thanks for noticing this reference, in this case, we refer to the citation of the Python implementation. We also included the citation for the methodology (Snoek et al., 2012). (see lines 263-264 in the re-revised version of the manuscript)

**Figures and Tables**

Fig. 1 there is a typo in the caption of ”Münsteländer Kreidebecken”. Also I would suggest to add the English names to the caption in addition to the legend so that readers do not have to speculate on the translations.

We agree to add the English names on the caption to improve clarity. Caption of Figure 1 of the re-revised manuscript has been modified to: “Hydrogeological areas of Lower Saxony. 1:500,000 (modified from LBEG (2016)). The hydrological bodies towards the north correspond to porous aquifers

(nord- und mitteldeutsches Mittelpleistozän (North-Central Middle Pleistocene), Niederungen im nord- unmitteldeutschen Lockergesteinsgebiet (North-Central lowlands in unconsolidated rock), Nordseemarchsen and Nordseeinseln und Watten (North Frisian Wadden sea, marsh islands and halligen)). The south consists of fractured and karst aquifers (Mitteldeutscher Buntsandstein (Central Bunter sandstone), Mitteldeutsches Grundgebirge (Central crystalline basement), Münsteländer Kreidebecken (Münsterland Chalk Basin), Nordwestdeutsches Bergland (Northwest Uplands), Sandmünsterland (Sand Münsterland) and Subherzyne Senke (Subhercynian Trough) ).”

Fig. 3 I miss a bit of information in the caption. What are the sources in a)-c)? Which year of CORINE? Especially given fractions calculated and described in Tab. 2 I am not sure what is shown here, what does “associated” mean here?

Thank you for your input. We agree that the caption is missing information regarding the sources. By "associated," we meant characteristics taken from the specific site. We modified the caption of Figure 3 for clarity to “The bar plots show the distribution of well characteristics in the study area: a. Aquifer type, b. Aquifer material (BGR, 2019b), c. Geomorphology (SMI: soil moisture index) (BGR, 2006), and d. CORINE Land Cover (Copernicus, 2018).”.

Tab. 1: As source, I would like to also provide the reference/link to the data, not only the provider, potentially also access date, version etc to increase the reproducibility

The groundwater observations and meteorological data used for this study are accessible through the complete citation of the corresponding source (see Gomez (2024) list of the re-revised paper).

Tab. 2: “up to … km“ do you mean that all distances larger than that were set to the limit?

Thanks for pointing this out. That is not the case; we meant that distances larger than the threshold were not included. For clarity, we modified it in text to “Distance to the coastline within 25 km” on Table 2.

Tab. 3: Stability – what is the window size? Fourier power spectral density - Why should it be annual climate variability, while it is actually groundwater variability?

Thanks for noticing this. For Stability, we have added a window size of 10 values to Table 3 in the re-revised manuscript. Regarding the Fourier power spectral density, we intended to estimate the annual periodic component of the GWL, as this might be due to the influenced by annual climatic factors. We modified the text “Higher values indicate a strong annual periodicity in GWL variability, which may be influenced by annual climatic factors.” on “*Implications for the GWL”* column on Table 3 in the re-revised manuscript.

**References**

Fritsch, F. N., & Butland, J. (1984). A Method for Constructing Local Monotone Piecewise Cubic Interpolants. SIAM Journal on Scientific and Statistical Computing, 5(2). https://doi.org/10.1137/0905021

Gomez, M. (2024). mgomezo12/Performance_CNN_v3: Assessing Groundwater Level Modelling using a 1D-CNN: Linking Model Perfor415 mances to Geospatial and Time Series Features (Version 3). Zenodo. https://doi.org/https://doi.org/10.5281/zenodo.12531372

Heudorfer, B., Liesch, T., & Broda, S. (2024). On the challenges of global entity-aware deep learning models for groundwater level prediction. Hydrology and Earth System Sciences, 28(3). https://doi.org/10.5194/hess-28-525-2024

Knoll, L., Breuer, L., & Bach, M. (2019). Large scale prediction of groundwater nitrate concentrations from spatial data using machine learning. Science of the Total Environment, 668. https://doi.org/10.1016/j.scitotenv.2019.03.045

Kratzert, F., Gauch, M., Klotz, D., & Nearing, G. (2024). HESS Opinions: Never train an LSTM on a single basin. Hydrol. Earth Syst. Sci. Discuss. [Preprint].

Knoll, L., Breuer, L., & Bach, M. (2019). Large scale prediction of groundwater nitrate concentrations from spatial data using machine learning. Science of the Total Environment, 668. https://doi.org/10.1016/j.scitotenv.2019.03.045

Tarasova, L., Gnann, S., Yang, S., Hartmann, A., & Wagener, T. (2024). Catchment characterization: Current descriptors, knowledge gaps and future opportunities. In Earth-Science Reviews (Vol. 252). https://doi.org/10.1016/j.earscirev.2024.104739

Wriedt, G., & NLWKN. (2020). Grundwasser Grundwasserbericht Niedersachsen Sonderausgabe zur Grundwasserstandssituation in den Trockenjahren 2018 und 2019. https://www.nlwkn.niedersachsen.de/download/156169/NLWKN_2020_Grundwasserbericht_Niedersachsen_Sonderausgabe_zur_Grundwasserstandssituation_in_den_Trockenjahren_2018_und_2019_Band_41_.pdf