Performance assessment of geospatial and time series features on groundwater level forecasting with deep learning.

Mariana Gomez[1,2], Maximilian Nölscher[2], Andreas Hartmann[1], and Stefan Broda[2]

[1]Institute of Groundwater Management, TU Dresden, Dresden, Germany
[2]Federal Institute for Geosciences and Natural Resources, Berlin, Germany

Correspondence: M.Gomez (mariana.gomez_ospina@tu-dresden.de)

Abstract. Groundwater level (GWL) forecasting with machine learning has been widely studied due to its generally accurate results and little input data requirements. Furthermore, machine learning models for this purpose are set up and trained in a short time when compared to the effort required for process-based numerical models. Despite the high performance of models obtained at specific locations, applying the same model architecture to multiple sites across a regional area might lead to contrasting accuracies. Likely causalities of this discrepancy in model performance have been barely examined in previous studies. Here, we investigate the link between model performance and the effects of geospatial site and time series features. Using precipitation (P) and temperature (T) as predictors, we model groundwater levels at approximately 500 observation wells in Lower Saxony, Germany, applying a 1-D convolutional neural network (CNN) with a fixed architecture and hyperparameters tuned for each time series individually. The GWL observations range from 21 to 71 years, leading to a variable test and training dataset time range. The performances are evaluated against relevant geospatial characteristics (e.g. landcover, distance to ~~water works~~waterworks, and leaf area index) and time series features (e.g. autocorrelation, flat spots, and number of peaks) using Pearson correlation coefficients. We found that model performance is negatively influenced at sites near waterworks and densely vegetated areas. Longer subsequences of GWL measurements above or below the mean negatively impact the metrics~~and might be associated with anthropogenic influence or wetter and drier periods~~. Besides, ~~complex~~ GWL time series containing more irregular patterns and with a higher number of peaks exhibit better metrics, possibly due to a closer link with precipitation dynamics. As deep learning models are known to be black-box models missing the ~~physical processes understanding~~understanding of physical processes, our work shows new insights into the degree of affectation that external physical factors have on the input-output relation of a GWL forecasting model.

keywords: Groundwater levels, deep learning, forecast

1 Introduction

Global water use increases, aggravated by climate change in current water-stressed areas and generating future stress in regions of abundant supply (UNESCO, 2022). Under these situations, groundwater is seen as a solution to ensure water supply, accounting for approximately 25% of the global freshwater (UNESCO, 2022). In fact, aquifers might report a seasonal and multi-year buffer capacity, but when deficits in groundwater storage are observed, they may last much longer due to the memory effect (UNESCO, 2020). ~~Investigating~~ Moreover, only slight groundwater level (GWL) ~~changes constitutes a way to estimate groundwater stress in the near and long term since only slight GWL~~ declines are needed to significantly affect groundwater discharges to streams (de Graaf et al., 2019). Consequently, approaches based on groundwater observation sites are valuable for estimating groundwater stress in the near and long term. This allows identifying ~~, among others,~~ over-exploitation based on depletion trends (Daliakopoulos et al., 2005), increasing the knowledge about water availability for drinking water supply and agricultural irrigation (Takafuji et al., 2019), and delineating potential soil subsidence zones due to extremely low groundwater levels in connection with droughts and water abstraction (Xu et al., 2008).

Physical and numerical approaches have been widely used as the primary tool to study GWL (Goderniaux et al., 2015). However, achieving a desired model calibration/validation requires extensive physical knowledge of the study area and large volumes of data related to the aquifer properties, geology, and topography, among others. In the last two decades, many publications have shown that data-driven models are simpler and faster to develop and provide more accurate results than physical or numerical models under certain conditions (Tao et al., 2022; Malik and Bhagwat, 2021; Ahmadi et al., 2022). Data-driven models using machine learning (ML) techniques such as artificial neural networks (ANNs) have proven their suitability for GWL forecasting (Wunsch et al., 2022) and the ability to capture the non-linearity of the aquifer's dynamics, although at the expense of having a physical understanding of the process. Many studies address the former challenge by applying explainable AI methods such as SHAP to elucidate the input-output non-linear dynamics (Chakraborty et al., 2021; Zhang et al., 2023; Liu et al., 2022). In particular, ~~ANNs~~ ANN are suitable for solving groundwater-related problems on a regional scale due to their low dependency on field data accessibility. Many ANN approaches have been successfully implemented, and recent developments in the field of deep learning (DL) promise a significant improvement of already existing prediction approaches. High overall performances have been obtained through ANNs techniques including feed-forward neural network (FFNN) (Roshni et al., 2020), long short-term memory (LSTM) (Wunsch et al., 2021), and convolutional neural networks (~~CNNs~~CNN) models (Mohanty et al., 2015;Ahmadi et al., 2022 ;Wunsch et al., 2022). Besides DL techniques, shallow recurrent networks such as non-linear auto-regressive networks with exogenous input (NARX) are proven to be useful for ~~modeling~~ modelling a wide variety of dynamic systems (~~Wunsch et al., 2018). In terms of~~ Guzman et al., 2017; Zanotti et al., 2019; Fabio et al., 2022). Regarding accuracy and calculation speed, the CNN models outperform the LSTM. NARX models performed, on average, better than CNN (Wunsch et al., 2021), mainly because NARX models capture temporal dependencies on groundwater. However, the ~~last one~~ CNN model has been shown to be faster with only a slightly lower accuracy (Wunsch et al., 2020).

Most studies have successfully applied these techniques for GWL forecasting using ~~only~~ meteorological variables as inputs. Up to date, the research focuses on a comparative analysis among different AI techniques, resulting in slight differences among models' performance (Wunsch et al., 2021) or in improving the model's accuracy by modifying its architecture (Gong et al., 2016). In many cases, disregarding site geospatial characteristics can reduce model accuracy or credibility, owing to the different responses depending on the aquifer characteristics (Kløve et al., 2013), unsaturated zone conditions, and groundwater contributing area (Rust et al., 2018). Therefore, it is ~~still known that~~ known that in order to achieve more accurate results in areas influenced by natural and anthropogenic factors~~;~~, river water level and human impact factors such as pumping rates should be considered as inputs (Lee et al., 2019). For instance, Gholizadeh et al. (2023) applied an LSTM model including static input features (e.g. hydraulic conductivity and soil depth) as an attempt to model ungauged locations, the authors attribute the satisfactory model performance to such inputs.

Since regional studies frequently lack supplementary information beyond meteorological data, this study explores the link between model performance (using only precipitation (P) and temperature (T) as inputs) vs. site-specific and time series features that might help to understand the input-output relation of a GWL DL model. Although many types of ANN structures have been developed for GWL forecasting, a 1-D CNN (LeCun et al., 2015) is applied here to evaluate the model performance due to their flexibility, calculation speed, and reliability. The model is trained, validated, and tuned individually in 505 wells distributed throughout the state of Lower Saxony, Germany. The research considers the relevant and available geospatial features and time series features. New insights are provided about the complexity of controlling factors on the groundwater dynamics.

2 Study area and materials

2.1 Study area

The study area is located in ~~the Lower Saxony(LS)~~Lower Saxony, Germany (Fig 1), where groundwater accounts for 86% of the public water supply (LSN, 2016). The groundwater bodies in this area comprise a great extension of highly productive porous aquifers and, in less proportion, fractured hard rock, and karst aquifers (LSN, 2016). The landscape is mainly dominated by the lowlands in the northern and central regions, whereas the south is predominantly hilly and mountainous. Land use corresponds mainly to farming (∼ 47%) and pasture (∼ 15%), concentrated in the western and northern regions (NMUEK, 2015). The maritime influence in the coastal region affects the precipitation distribution, decreasing from the West (approx. 750mm/yr) to the East (<600 mm/yr). In contrast, the annual precipitation exceeds 1500 mm in the south (NMUEK, 2015).

From a broad perspective, the northern German Plain is covered up to the edge of the low mountain range by glacial deposits of varying thicknesses (LBEG, 2016), constituting a great proportion of ~~LS~~Lower Saxony. Hard rock areas in the southern highlands are formed by sandstones and limestones (BGR, 2019). Highly heterogeneous geological structures exist among these two groups, leading to groundwater availability at different depths with varying yields, especially in karst aquifers (LBEG, 2016). The primary pressures on the quantitative status of groundwater bodies arise from its long-term abstraction, mainly for drinking water, irrigation, mining or construction activities, and long-term hydraulic measures for groundwater remediation (NMUEK, 2015).

2.2 Data

GWL observations and meteorological information are available throughout the state of ~~LS~~Lower Saxony. Table 1 shows the data overview. The GWL is in monthly resolution with a variable time range, and historical records of meteorological variables are available in a daily resolution of 5 x 5 km. The GWL time series consists of 505 wells that are unevenly distributed, with more information available in the central region of the study area~~(Fig. 2).~~ . Besides the irregular spatial distribution, ~~the~~ there are data gaps depending on the well (Fig. 2.a), and the time range of the groundwater records ~~is highly variable (between 1950~~

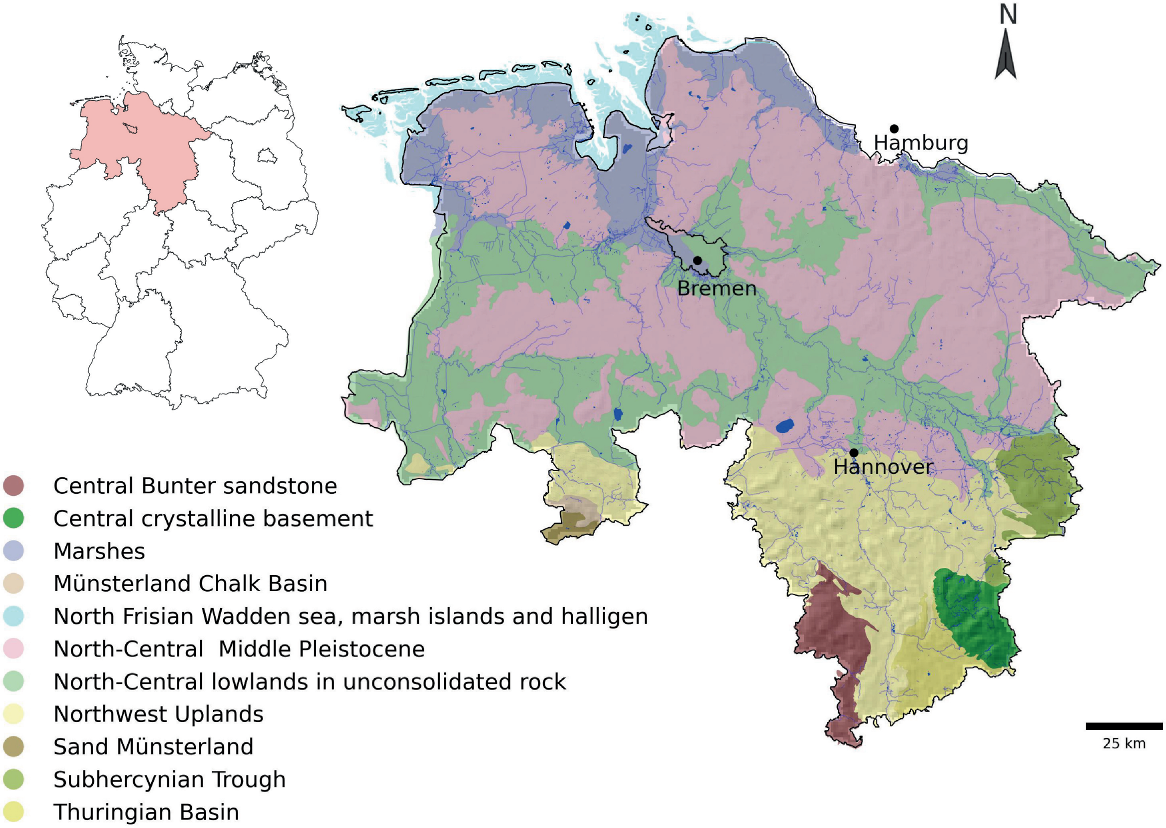
N
Hamburg
Bremen
Hannover
25 km
● Central Bunter sandstone
● Central crystalline basement
● Marshes
● Münsterland Chalk Basin
● North Frisian Wadden sea, marsh islands and halligen
● North-Central Middle Pleistocene
● North-Central lowlands in unconsolidated rock
● Northwest Uplands
● Sand Münsterland
● Subhercynian Trough
● Thuringian Basin

Figure 1. Hydrogeological areas of Lower Saxony. 1:~~500K~~ 500,000 (~~LBEG, 2016~~modified from LBEG (2016)) . The hydrological bodies towards the north correspond to porous aquifers (Nord- und mitteldeutsches Mittelpleistozän, Niederungen im nord- un mitteldeutschen Lockergesteinsgebiet, Nordseemarchsen and Nordseeinseln und Watten). The south consists of fractured and karst aquifers (Mitteldeutscher Buntsandstein, Mitteldeutsches Grundgebirge, Münsteländer Kreidebecken, Nordwestdeutsches Bergland, Sandmünsterland and Subherzyne Senke)

~~and 2021) , and considerable data gaps exist depending on the well~~varies between 21 and 71 (Fig. 2.b) years from 1950 to 2021, resulting in differences in start-end dates of time series.

Table 1. Data availability overview.

Data	Temporal resolution	Spatial resolution	Time range	Source
Groundwater level observations	Monthly	-	Variable (1950 : 2021)	The Lower Saxony State Office for Mining, Energy and Geology (LBEG)
Precipitation and temperature	Daily	5 x 5 km	1951 : 2015	(Rauthe et al., 2013; Frick et al., 2014)

~~The uneven spatial distribution of the wells results in fewer data available in those areas with fractured aquifers (~~As observed in Fig. 3.a~~),~~ , less data is available for fractured aquifers, limiting the interpretation in terms of different hydrogeological units. This uneven spatial distribution of the wells reflects the differences in hydraulic properties between porous and fractured aquifers. In the latter, water primarily flows through conduits and cavities, creating a more complex system that could increase the construction and maintenance costs of wells, reducing their number in the area. Almost half of the wells are located in sandy-gravel material (Fig. 3.b), associated with high hydraulic conductivity and stronger variations of GWL. The other half is in finer materials but still with a high sand portion. Regarding geomorphology, the predominant category is low relief with a high to moderate soil moisture index (SMI), followed by sink areas with a high SMI (Fig. 3.c). Most wells are in non-irrigated arable lands and pastures (Fig. 3.d). Overall, the study area characteristics associated with each well are relatively homogeneous regarding hydrogeology, geomorphology, and land use. Most wells are located below 100 m.a.s.l. (northern area), and higher elevations relate to wells in the southern mountainous regions. According to the filter depth, most analyzed wells relate to shallow aquifers (Fig A1).

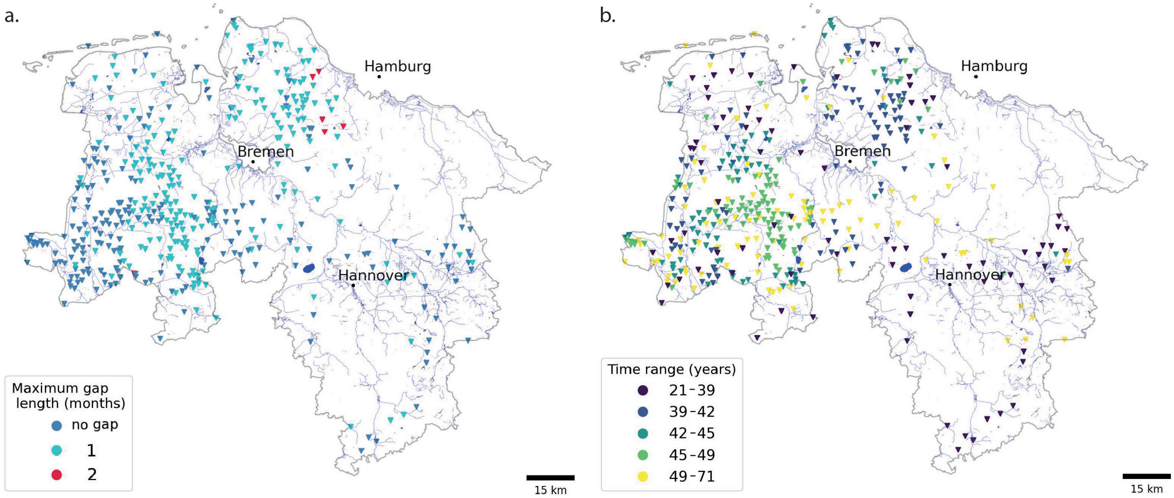

Figure 2. Location of the 505 wells with GWL time series observations used in the study. a. Maximum gap length and ~~time~~ b. Time range of the GWL time series. Author-generated map.

The historical records of meteorological information in Germany are available as an observational dataset (HYRAS dataset, Rauthe et al. (2013), Frick et al. (2014)). This corresponds to gridded hydrometeorological information based on a compilation of variables across Germany and adjacent river basins (Razafimaharo et al., 2020). The dataset consists of daily precipitation (interpolated according to Rauthe et al. (2013)) and temperature from ~~1951-2015.~~ 1951 to 2015. The German Weather Service (DWD) adapted and improved the raster data based on more than 1300 stations and with a direct station-grid comparison, making the data highly reliable (Razafimaharo et al., 2020). The daily dataset is provided free of charge for academic and non-commercial purposes ~~by the DWD~~(DWD, n.d.).

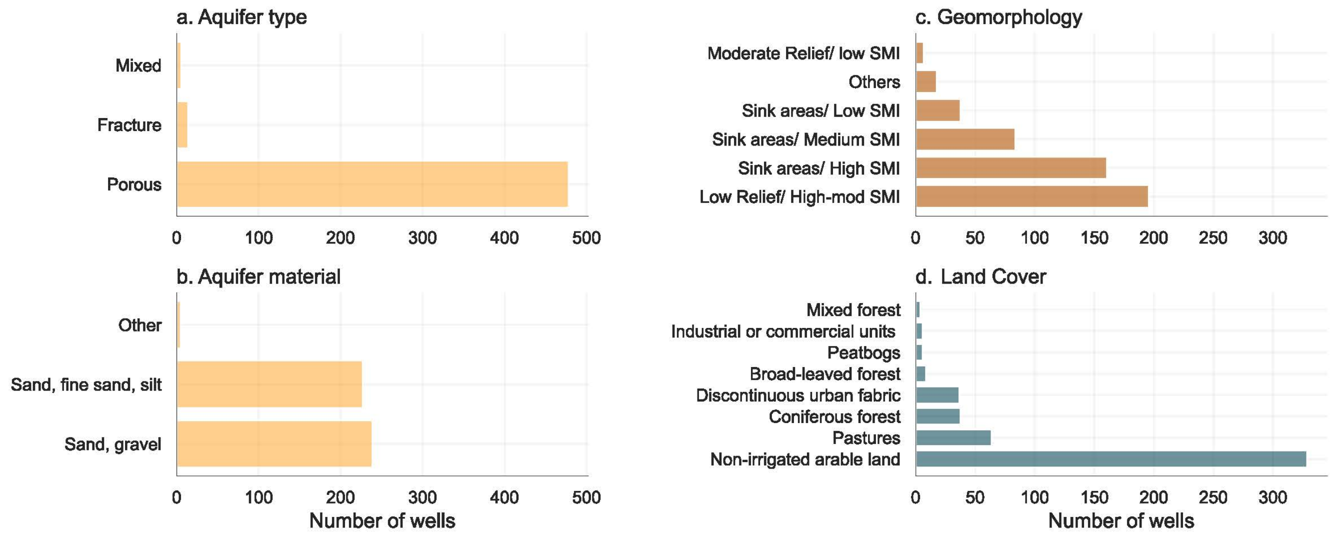

Figure 3. Bar plots showing ~~(a) aquifer~~ . Aquifer type, ~~(b) aquifer~~ . Aquifer material, ~~(~~c. Geomorphology (SMI: soil moisture index) ~~geomorphology~~, and ~~(d)~~ . land cover (CORINE Land Cover) associated with the studied wells.

3 Methods

Figure 4 presents the methodological flow chart. the first stage consists of pre-processing the available information, jointly with exploratory data analysis and data mining. The procedure starts with the GWL observations involving the filtering, data imputation, and jump detection steps. Simultaneously, the meteorological variables are extracted per well location and re-sampled to monthly resolution. As a result, there is an input dataset per well relating GWL, P and T. In the second stage, a CNN model is implemented, validated, optimized, and tuned through a Bayesian optimization process. The latter corresponds to an optimization method based on bayesian inference and Gaussian process to maximize the sum of performance metrics, in this case NSE and R^2. The following step is the performance evaluation and interpretability, relating geospatial and time series features with the performance metrics ~~. The final discussion intends to link the results from the input data analysis, interpretability, and forecasting steps.~~ (Snoek et al., 2012; Fernando Nogueira, 2014) . To achieve the objectives, several Python libraries are used: Pandas (Reback, 2020), Numpy (Van Der Walt et al., 2011), Scipy (Virtanen et al., 2020), Matplotlib (Hunter, 2007), Geopandas (Jordahl et al., 2020), and Tensorflow (Abadi et al., 2015) as the most relevant throughout the process. Additional specific libraries are later mentioned at each methodological step.

3.1 Preprocessing

The ~~initial~~ initially available GWL information consists of 962 wells, but a selection is ~~done~~ made to exclude wells under strong anthropogenic influences such as pumping, ~~favoring~~ favouring the dependency between the ~~climatic input variables and the groundwater observations~~meteorological input features and observed groundwater levels. After applying this ~~criterion~~filter,

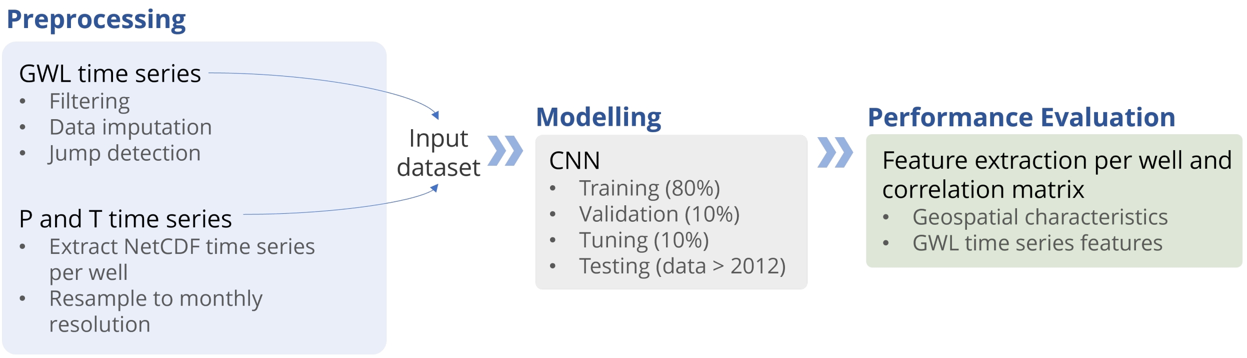

Figure 4. Methodological flow chart.

a total of 745 wells remain. A second selection removes time series with gap lengths above ~~three months (2~~ two consecutive missing values~~)~~, obtaining 505 wells, 241 (48%) as a complete series, 254 ~~with two months as the maximum gap~~(50%) with one missing value, and 10 ~~with three months gap~~(2%) with two missing values. 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) and~~ . This method is only applied when the wells exhibit similar behaviour in their time series, as measured by the Euclidean distance. otherwise, we use the Piecewise Cubic Hermite Interpolating Polynomial (PCHIP) ~~.~~ for gap filling. Overall, the time series have less than 5% gap-filled values. Additionally, jumps (sudden changes in the time series) are ~~present at some~~ identified at 28 wells and might be associated with measurement instruments or other technical problems (Post and von Asmuth, 2013; Retike et al., 2022). We ~~removed~~ identified the observations displaying these anomalies by finding the highest slope in the cumulative sum ~~.~~ and removing the time series before 1990 for those wells. This is because we are aware of changes due to measurement devices around this time. Finally, to extract the meteorological information, an average of 3 x 3 pixels is used to reduce uncertainty related to the grid cell size following the suggestion of Linke (2017).

3.2 Modelling

The ~~1d-CNN~~ 1D-CNN structure is implemented based on Wunsch et al. (2022), where inputs are split into sequences of a defined value given by the sequence length. The sequences of input data pass through a 1D convolutional layer, where a window of a fixed kernel convolves through the data. The maximum of each convolution is extracted to generate the max pooling layer. A Monte Carlo dropout of 50% is made to avoid model overfitting. This is followed by a flattened layer and a fully connected dense layer that uses a rectified linear unit (RELU) as the activation function.

The CNN model is applied to each GWL time series, and ~~,~~consequently, the phases of training, validation, optimization, and hyperparameter tuning are also carried out per well. The available groundwater data before 2012 is split between the training (80%), validation (10%), and hyperparameter tuning (10%) dataset, and the ~~time-series~~ time series after 2012 is used as the test set. Each subset differs depending on the time range of GWL observations (ranging from 21 to 71 years). Therefore, the input features, time range, and specific model parameters make the model a unique representation of the GWL in a particular location. An Adam optimizer is applied with 100 training epochs, an initial learning rate of 0.001, and the early stopping of 15 patience. In this case, the loss is minimized with the mean squared error (MSE) through each epoch for the validation process. The hyperparameter tuning is done with a Bayesian optimization (Fernando Nogueira, 2014) to maximize the sum of the squared Pearson (R^2) ~~(eq. ??)~~ and the Nash-Sutcliffe efficiency (NSE) ~~(eq.??)~~ coefficients, measuring the deviation of observed ~~(obs) from predicted ($pred$)~~ from predicted GWL over a total of ~~n~~ observations. The hyperparameters correspond to: kernel size (fixed as 3), sequence length (1-12 months), number of filters (1-256), dense size (1-256), and batch size (1-256). Owing to the dataset's monthly resolution, the sequence length boundaries are set between 1 and 12 months, a time range that can include significant variabilities in the sub-sequences.

$$R^2 = \left[\frac{\sum_{i=1}^{n}(Y_i^{obs}-\overline{Y}^{obs})(Y_i^{pred}-\overline{Y}^{pred})}{\sqrt{\sum_{i=1}^{n}(Y_i^{obs}-\overline{Y}^{obs})^2(Y_i^{pred}-\overline{Y}^{pred})^2}}\right]^2$$

$$NSE = 1-\frac{\sum_{i=1}^{n}(Y_i^{obs}-Y_i^{pred})^2}{\sum_{i=1}^{n}(Y_i^{obs}-Y_i^{mean})^2}$$

3.3 Performance evaluation

The model performance ~~is influenced to a significant or minor degree~~ can be significantly or slightly affected, depending on the well location, by natural and anthropogenic factors, such as the distance to waterworks or watercourses, the type of land cover, and the geomorphology. Besides, the intrinsic patterns present in the observation time series might reveal external affectations on the GWL model. Table 2 describes the geospatial features considered. ~~Among them~~We also performed the analysis with further geospatial features, such as distance to the surface water bodies, but no statistically significant correlation with model performance was found, and therefore, the results are not shown here. Among the reported ones, the distance to the waterworks is expected to modify groundwater flow and, consequently, the GWL nearby in the surrounding wells. ~~In this case, it is assumed that the date extracted from~~ Here, we assume that Open Street Map (OSM, 2022) includes a significant proportion of all waterworks in the study area~~. Still, direct influences are unknown due to the lack of pumping rates information~~, but a comprehensive dataset including the locations of all waterworks or information regarding pumping rates is still missing. Regarding categorical variables, the proportion of a 1 km ~~buffer surrounding~~ radius around the well is taken for the most relevant categories. The Python packages of Tsfeatures (Yang and Hyndman, 2020) and Tsfresh (Christ et al., 2018) are used to extract multiple GWL time series features automatically. A selection is made ~~among~~ from the long list of features (available in each package) according to their correlation coefficient in relation to the metrics and the added value to the analysis. ~~An~~ Table 3 shows an overview of the selected ~~features is given in Table 3, and~~ time series features, description, range of values and guidelines of their occurrence on the GWL time series (for a detailed description of the estimation procedure, please refer to the package manual~~. Sample and approximate entropy, Fourier entropy, Lempel Ziv complexity, and the number of peaks are features to estimate the time series complexity.~~

~~To evaluate the impact of external factors on the model performance, the geospatial and time seriesfeatures are extracted per well and correlated with the accuracy metrics (R^2, NSE, and BIAS) through the Pearson correlation coefficient. An R^2 and NSE value closer to~~). We incorporated the Fourier power spectral density at a period of 1 ~~mean a higher similarity between modeled and observed GWL, whereas the closest the BIAS is to zero, the more similar are simulations to the results, negative BIAS refers to a model with underestimation. To enhance the robustness of the correlations, we took the mean correlation coefficient after bootstrap sampling with 100 re-sampling datasets. Only statistically significant correlations within a 90% confidence interval are reported. The main objective is to notice positive or negative effects on the model performance.~~ year to measure the influence of annual climate seasonality on the GWL. Higher values indicate a greater annual seasonality. High autocorrelation values indicate patterns constantly repeating in the time series. High stability values imply that GWL remains within a consistent range without significant variations or trends. The more flat spots, the more relatively constant values over extended periods. Approximate entropy and number of peaks measure the complexity of the time series. A high value of the former indicates that the GWL time series contains multiple irregular patterns, making it harder to predict. A higher number of peaks indicates multiple local maximums, implying stronger fluctuations in GWL observations.

To evaluate the impact of external factors on the model performance, the geospatial and time series features are extracted per well and correlated with the accuracy metrics (R^2, NSE, and bias) through the Pearson correlation coefficient. An R^2 and NSE value closer to 1 mean a higher similarity between modelled and observed GWL, whereas the closer the bias is to zero, the more similar are simulations to the observed data; negative bias refers to a model with underestimation. To enhance the robustness of the correlations, we took the mean correlation coefficient after bootstrap sampling with 100 re-sampling datasets. Only statistically significant correlations within a 90% confidence interval are reported. The main objective is to notice positive or negative effects on the model performance.

4 Results

4.1 Modelling

The performance per well is presented in Figure 5~~, together with the histogram~~. According to our results, a total of 212 wells show R^2 and NSE values above 0.7 and 0.6, respectively (Fig. 5), which we would consider an acceptable model ~~quality~~ fit (Moriasi et al., 2015). Lower performance is seen mainly in the south, related to the fractured aquifers, where both metrics (R^2 and NSE) are below 0.5. The highest positive and negative ~~Bias~~ bias also occurs in those hydrogeological areas. These wells correspond to the shortest data length. Most of the best-performed models are found ~~in~~ for the wells in the central region of the study area, where the density of wells is higher. Contrarily, ~~near the coast,~~ some models exhibit low performance near the coast regarding R^2 and NSE, ~~but Bias is in~~ with a bias is between ± 0.2.

After a visual comparison of most of the CNN models and GWL observations, an overall good agreement is visible between the simulated GWL and the observations. Figure 6 shows examples where the optimized model performs well and where the model does not correctly ~~reproduces~~ reproduce GWL variability. As observed, the model sometimes underestimates and overestimates the peaks and lows. However, steep peaks are mainly underestimated. In most cases, local variations ~~out of~~

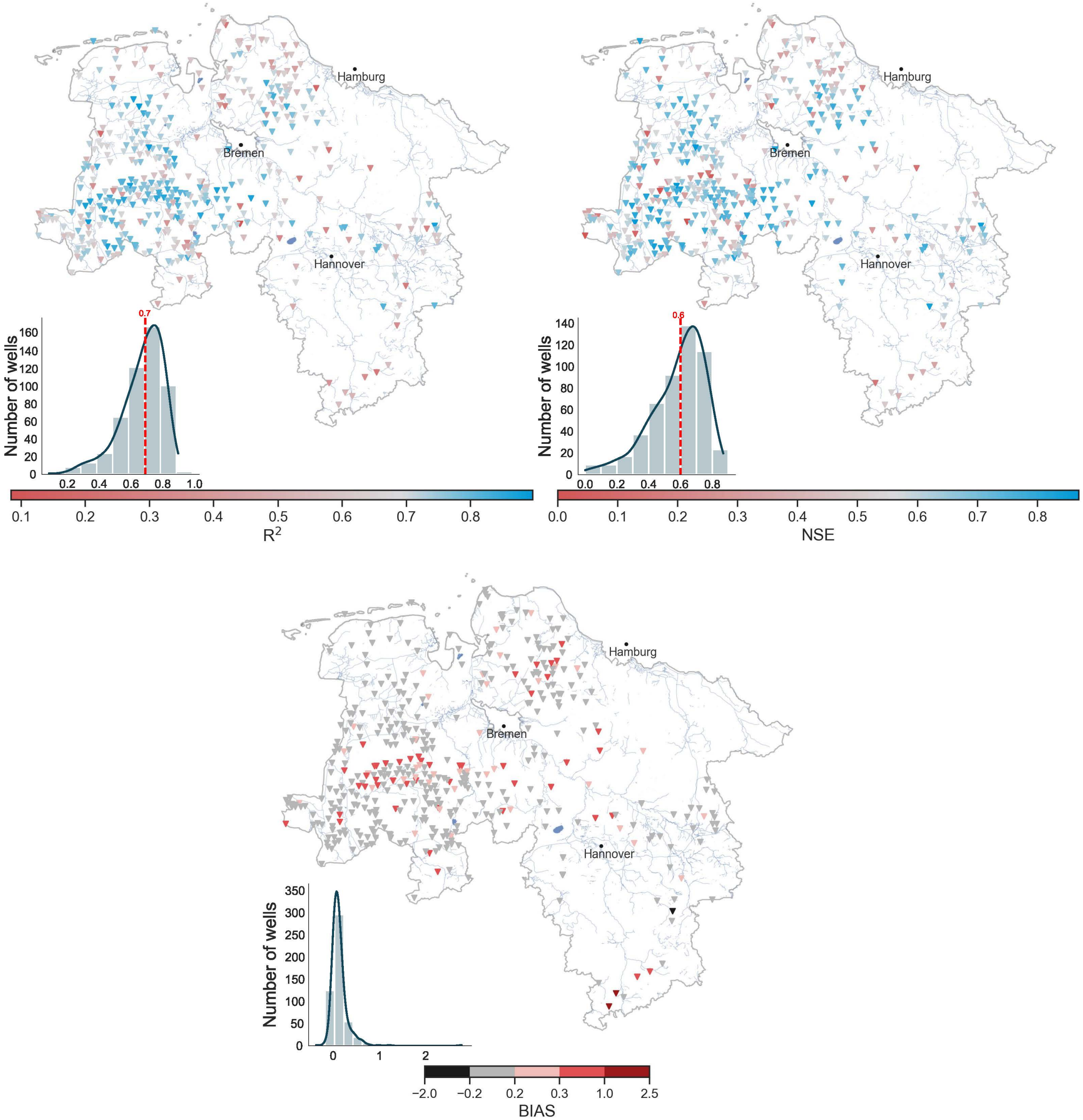

Figure 5. Spatial distribution of model performance metrics (R^2, NSE and Bias) per well and their respective histogram. Author-generated map.

from the main seasonal ~~behavior~~ behaviour are ignored. Occasionally, in poorly performing models, the pattern of the GWL observations has been generally learned but with a strong ~~Bias~~bias (around 10% of the wells show a bias above 0.13). The well-performed cases show how the CNN model can represent low peaks for some wells.

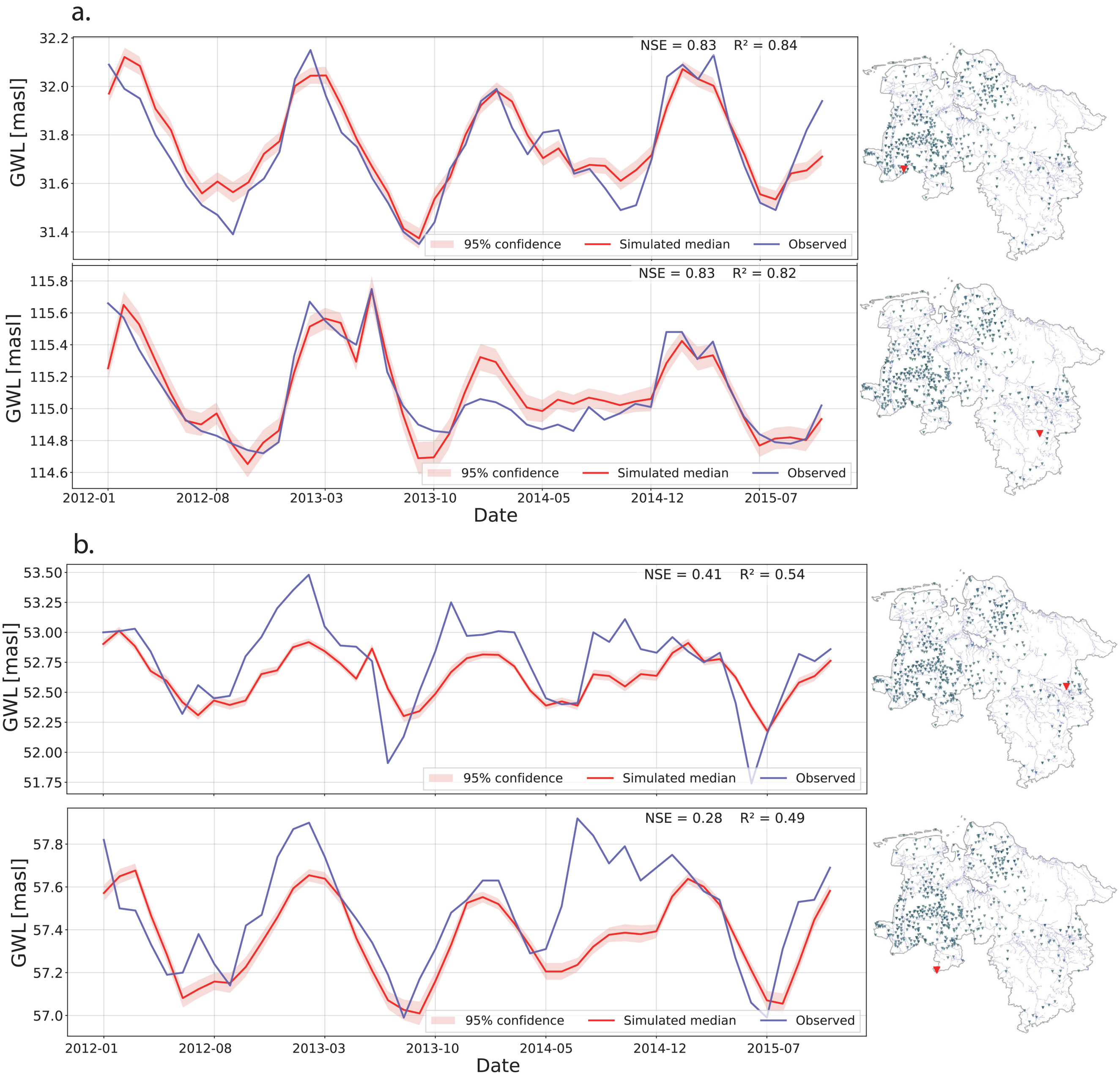

Figure 6. Examples of the optimized model with (a) high performance and (b) low performance.

4.2 Performance assessment

The correlation coefficients between the geospatial, and time series features and the model performance are shown in Figure 7. Only significant correlations with a confidence interval level of 90% are displayed. Although correlation coefficients are statistically significant, they do not exceed 0.62. One of the highest correlations is the distance to the waterworks, corresponding to 0.43 (R^2) and 0.29 (NSE). The R^2 increases as the distance to the coastline does, whereas the ~~Bias~~ bias reduces. The proportion of the most common landcover 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 ~~metrics values~~correlations. Sink and low relief areas with medium to high SMI negatively impact performance. Hilly regions evidence lower accuracy, while areas with a high drainage density or topographic wetness index evidence a ~~better fit to the simulated GWL~~higher model performance.

Stronger correlations, mainly negative, are found for the time series features. Overall, we found that autocorrelation reduces model performance~~, which~~. This might not be the case when using antecedent GWL as an additional input feature when GWL will show 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 mean variance over overlapping windows) ~~, seasonal trend, and strength reduce~~ reduces the model performance. Increasing flat spots and long strikes above or below the mean are negatively correlated, mainly concerning the NSE metric. The positive correlations are mainly associated with the complexity measures such as ~~sample and approximate entropy , Lempel Ziv complexity,~~ approximate entropy and the number of peaks. The time series length positively correlates with R^2 but does not correlate with NSE. Stronger values of the Fourier power spectral density at one year (stronger annual seasonality on the observed GWL) result in a higher model performance.

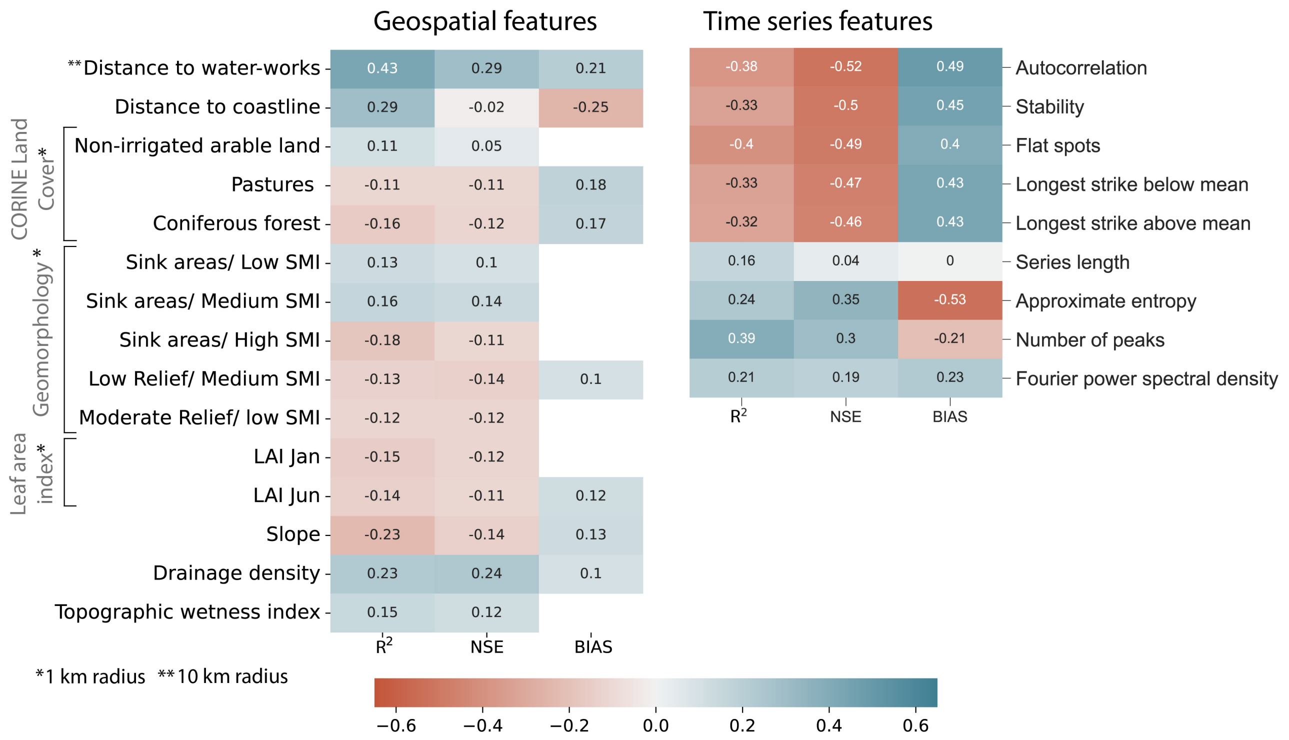

Figure 7. Pearson correlation coefficients between the geospatial features, GWL time series features, and the model performance. Significant correlations are displayed with a confidence level of 90%. Blank spaces correspond to non-significant correlations. Correlations with the distance to waterworks are done with 90 wells located in the 10 km buffer and with 50 wells located up to 25 km for the distance to the coastline.

5 Discussion

The analyzed wells are located in a relatively homogeneous area in terms of hydrogeology, associated with a major proportion of porous material and shallow aquifers, improving the model's capacity to express GWL only in terms of meteorological inputs (Kløve et al., 2013). There are a few wells in the fractured and karst aquifers, but those are frequently associated with greater depths (Wunsch et al., 2022). A more diverse distribution ~~is seen regarding the~~ of wells is observed regarding land cover and geomorphology~~. The difference in land use associated with each well implies a complex system that interacts with climate variability and affects the groundwater resources~~ , resulting in distinct interactions between climate, land use, and groundwater (Kløve et al., 2013;Treidel et al., 2011), potentially influencing the model performance.

The primary source of uncertainty in the current analysis lies in the inability to separate the effects of each external ~~feature~~ features affecting observations (especially the geospatial features), which will greatly depend on the aquifer size (Kløve et al., 2013), amount of information available, and their reliability. To better interpret the non-linear behaviour between groundwater and its 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. Furthermore, since the magnitude of fluctuations of GWL varies greatly from season to season (Taylor and Alley, 2001), depending on the aquifer properties, groundwater dynamics are better observed at a weekly or even daily temporal resolution instead of the monthly time step. In addition, owing to the fact that the vast majority of the wells we used in the current analysis are located in porous aquifers, our results are mainly representative of these conditions.

The GWL behaves following the interaction between climate, topography, hydrogeology, and land use, among others (Earman and Dettinger, 2011). Estimating GWL solely with meteorological variables brings uncertainty, especially in areas with more significant human impact. Additionally, there are uncertainties related to the model realizations, which, in this case, are solved by using several random initialization seeds. ~~The uncertainties related to~~ As a result, the model precision ~~are reduced when using only~~ is generally high, and we only use the best-performed optimized models. Regarding the geospatial relations with the model performance, there are uncertainties based on the variable scale and the definition of influential ~~ratio~~ radius (assumed as 1 km for the geomorphology and land use, 10 km for the waterworks) and with the reliability of the primary information.

5.1 Modelling

Overall, the CNN model was able to simulate, to a significant extent, the GWL changes for more than 200 wells with good overall performance (R^2> 0.7 and NSE > 0.6). Thus, the remaining wells account for at least one metric with a non-acceptable performance, and in those cases, further hydrological or anthropogenic factors might influence the GWL ~~behavior~~behaviour. The Bayesian optimization currently maximizes the sum of R^2 and NSE, occasionally causing contrasting values for both metrics at specific wells. Thus, constraining both values to define model performance guarantees adequate results, even when individual accuracy is lower than acceptable criteria (Gong et al., 2016). Different combinations of metrics can also be explored against model improvements. As explained, Bayesian ~~optimization is~~ inference and Gaussian process (Fernando Nogueira, 2014) are used to tune the hyperparameters (external parameters that can not be learned from the data). However, additional tuning strategies such as Genetic Algorithm and Grid Search have shown better results (Alibrahim and Ludwig, 2021). Therefore, modifying the optimization strategy and following the standard approach of changing the network architecture can enhance the results. Other networks, such as LSTM or FFNN, can potentially increase the learning process. However, in the current study, understanding the influence of geospatial and temporal features related to the GWL has priority over network architecture.

Generalizing the model inputs for all wells throughout the state influences the scores, especially at sites where GWL is not only driven by P and T. Even with a low performance, sometimes the model can learn the GWL variations but incorporates a ~~Bias.~~ bias. 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 P and T) and the GWL observations can cause this bias at some stations. When both metrics used for the optimization (R^2 and NSE) are high, the model is seen to fit the observations adequately. At certain times, the model misses the small spikes on the observations. However, a model that adequately represents the lower and higher periods due to dry or wet years holds higher relevance for groundwater management. A low performance occurs mainly when a notorious anthropogenic or non-periodic signal is observed in the time series. Every model that could not correctly learn from meteorological inputs might be treated independently. Specific external forcings influencing GWL variability might be studied, and particular cases should be re-trained with the additional influencing variables.

5.2 Performance evaluation

The weak correlations between the geospatial features and the model performance can be related to the regional scale of the analysis and to the multiple drivers controlling the GWL at a specific location. Factors such as the spatial resolution of the geospatial features or the large numbers of observation pairs could also reduce the correlation coefficients (Armstrong, 2019). For instance, skewed probability distribution in the filter depth, which ~~in most wells~~ is below 50 m in most wells, excludes deeper aquifers from the analysis and can hinder the relation. Even though we found a directly proportional relationship between model performance and distance to waterworks, the correlation might be weaker due to non-reported abstractions. However, it is inferred that wells outside the influence area of the waterworks are more prone to be represented only by meteorological variables. Contrarily, wells located in the influence area of the waterworks system should include variables such as abstraction rates to keep the learning process stable (Lee et al., 2019)

The land cover can influence the recharge and the GWL dynamics. When the surface is sealed, the aquifer recharge decreases, and the GWL diminishes. In the same way, groundwater recharge is significantly reduced through evapotranspiration wherever dense vegetation is present, such as in a native forest (Lerner and Harris, 2009). In this case, most wells are located in non-irrigated arable land, which consists of rainfed crops, meaning a more direct response of GWL to meteorological variables is feasible. Indeed, as seen in Figure 7, the correlation is positive when the surrounding area of the well relates to a high proportion of non-irrigated arable land. Contrarily, model performance reduces as LAI increases. LAI indicates the vegetation canopy, and therefore, it governs the interception of precipitation, largely controlling the partitioning of infiltrated water into evapotranspiration and percolation (Reichenau et al., 2016). Thus, the interception process can hamper a direct response of GWL to precipitation (Pan et al., 2011), then affecting model performance. Regarding geomorphology, areas of accumulation (sink areas) with low to medium SMI positively affect the performance but negatively when the SMI is high. Sites with higher relief and SMI present lower performance. According to Rajaveni et al. (2017), geomorphological features referring to the accumulation process (pediment and valley fill) have a good groundwater potential and are, therefore, more prone to react to meteorological inputs. Accumulation areas are also represented by risen drainage density and TWI because these areas are feasible to respond quicker to meteorological inputs. We also expected the model's fitness to decrease as the slope increases since steeper areas account for higher runoff, reducing precipitation dynamics' influence on GWL observations.

As the geospatial characteristics surrounding the groundwater well influence observations, investigating the ~~intrinsic time series patterns reveals external affectationson the~~ patterns encountered in the time series by extracting selected features can provide insights into model performance affectations. For instance, the recurrent presence of flat spots on the observations, seen as relatively constant values over extended periods, reduces model performance. ~~We~~ This might indicate an aquifer that is less responsive to changes in climate, which is often the case with large aquifers (Kløve et al., 2013). We can apply a similar argument to the reduction of performance when there is an increase in time series stability. This means the GWL remains within a specific range of values without significant variations. Thus, even if there are upward or downward trends in precipitation, the observations of GWL do not exhibit similar patterns. Consequently, the proposed model using only P and T would fail to reproduce the GWL patterns adequately. We found that the learning process reduces as long consecutive subsequences above or below the mean occur. Direct human influences such as managed aquifer recharge can keep the GWL above the average and modify its response to meteorological variables. The opposite happens when ~~pumping occurs~~groundwater abstractions exceed recharge, and the aquifer levels drop for a more or less continuous period (Wendt et al., 2020). In both situations, the anthropogenic effects on GWL reduce the performance. Natural climate variability might also result in a similar effect, negatively affecting performance. For instance, if wetter or drier periods occur during testing but not in the training phase, the model is unlikely to learn the consequent patterns. Additionally, the time series complexity measures (~~sample and approximate entropy , Lempel Ziv complexity,~~ approximate entropy and the number of peaks) ~~evidence~~ indicate a directly proportional relationship with model performance, meaning that the more complex the GWL time series is (more irregular patterns), the better fit simulations with observations. Complex GWL time series might reflect a good response to precipitation.

Previous studies have shown little or no correlation between the time series length and the model performance (Wunsch et al., 2020). However, at least observations over decades are required to cover groundwater dynamics due to climate variability (Taylor and Alley, 2001), especially when considering a monthly temporal resolution. In this sense, the model can incorporate more information into the learning process, and model performance might increase with longer time series. However, conclusions about this relation should be further studied.

6 Conclusions

Fluctuations in the GWL observations are influenced by a combination of natural and anthropogenic factors, challenging the ~~modeling~~ modelling of groundwater systems. An alternative to high data-required physical and numerical models is DL techniques. Many DL models have been applied to GWL ~~modeling~~modelling, but the main concern about using these models remains a lack of physical understanding. Owing to the complex system between climate, GWL, and external drivers, model performance can be directly or indirectly affected outside of what the model can explain, limited by the input features. Our study brings insights into how model performance is affected by geospatial features and intrinsic time series characteristics. We selected a 1d-CNN model to simulate monthly GWL time series per well in northern Germany, using P and T as inputs. We found low performances in wells ~~nearby~~ near waterworks, an expected result as GWL are modified by pumping rates. An increased LAI or forest land cover leads to lower performance by hindering the P and T relation with the GWL. Complex time series show a better performance, possibly linked to a closer relationship between GWL and P patterns. More extended continuous GWL measurements above or below the mean negatively impact the metrics and can be associated with artificial recharge, pumping imposed in the time series, or natural events such as wetter and drier seasons. Even though only P and T are used as model inputs, the performances obtained are considered acceptable ($R^2 > 0.7$ and NSE > 0.6) for more than 200 wells.

As the study covers ~~are regional area~~regional areas, local variabilities in climate and human-water interactions might occur. Therefore, model performance should be evaluated at locations with greater data availability to strengthen the current research. Moreover, correlations might vary depending on the model architecture selected or the temporal resolution of GWL observations. For instance, daily resolution can better include groundwater dynamics showing stronger correlations. Our results encourage the joint analysis of physical-related characteristics and DL GWL ~~modeling~~ modelling as an essential path to improve the reliability of data-driven models.

Code availability. The code necessary to reproduce our results is available on GitHub under Gomez, M. (2024). Performance CNN. Zenodo. https://zenodo.org/doi/10.5281/zenodo.10625143

Data availability. All groundwater level data are available free of charge from the respective websites of the local authorities.

Appendix A: A

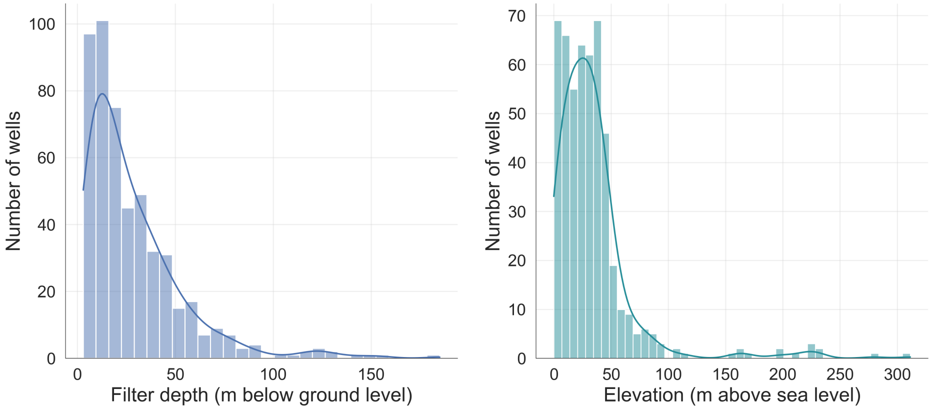

Figure A1. Filter depth (meters below ground level) and elevation in meters above sea level of all the wells in the study area.

Author contributions. Mariana Gomez: Methodology, Visualization, Data curation, Writing – original draft. Maximilian Nölscher: Conceptualization, Methodology, Writing – review, and editing. Andreas Hartmann: Supervision, review, and editing. Stefan Broda: Supervision, review, and editing.

Acknowledgements. This work was supported by the Erasmus Mundus scholarship following the Joint master's degree Program on Groundwater and Global Change – Impacts and Adaptation, and the FOSTER program of Technische Universität Dresden.

The Article Processing Charge (APC) was funded by the joint publication funds of the TU Dresden, including Carl Gustav Carus Faculty of Medicine, and the SLUB Dresden as well as the Open Access Publication Funding of the DFG.

Competing interests. No competing interests are present

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G., & Davis, A. (2015). TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://doi.org/10.5281/zenodo.4724125

Ahmadi, A., Olyaei, M., Heydari, Z., Emami, M., Zeynolabedin, A., Ghomlaghi, A., Daccache, A., Fogg, G. E., & Sadegh, M. (2022). Groundwater Level Modeling with Machine Learning : A Systematic Review and Meta-Analysis, 1–22.

Alibrahim, H., & Ludwig, S. A. (2021). Hyperparameter Optimization: Comparing Genetic Algorithm against Grid Search and Bayesian Optimization. *2021 IEEE Congress on Evolutionary Computation, CEC 2021 - Proceedings*. https://doi.org/10.1109/CEC45853.2021.9504761

Armstrong, R. A. (2019). Should Pearson's correlation coefficient be avoided? *Ophthalmic and Physiological Optics*, *39*(5), 316–327. https://doi.org/10.1111/opo.12636

Beven, K. J., & Kirkby, M. J. (1979). A physically based, variable contributing area model of basin hydrology. *Hydrological Sciences Bulletin*, *24*(1), 43–69. https://doi.org/10.1080/02626667909491834

BGR. (2006). GMK1000R Version 2.0. https://numis.niedersachsen.de/kartendienste;jsessionid=6B72740599919B86FCB8577E5249F31B?lang=de&topic=naturlandschaft&bgLayer=maps_omniscale_net_osm_webmercator_1&E=1013007.37&N=6912886.50&zoom=7&layers=2254e4d9c5a743729f1e886a188ec461&layers_visibility=

BGR. (2019). Geologische Übersichtskarte der Bundesrepublik Deutschland 1:250.000 (GÜK250). https://produktcenter.bgr.de/terraCatalog/DetailResult.do?fileIdentifier=0f2e1b5b-fc02-4491-a12b-2178473f5c84

BKG. (2021). Digitales Geländemodell Gitterweite 1000 m (DGM1000). https://gdz.bkg.bund.de/index.php/default/digitale-geodaten/digitale-gelandemodelle/digitales-gelandemodell-gitterweite-1000-m-dgm1000.html

Chakraborty, D., Başağaoğlu, H., Gutierrez, L., & Mirchi, A. (2021). Explainable AI reveals new hydroclimatic insights for ecosystem-centric groundwater management. *Environmental Research Letters*, *16*(11). https://doi.org/10.1088/1748-9326/ac2fde

Christ, M., Braun, N., Neuffer, J., & Kempa-Liehr, A. W. (2018). Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). *Neurocomputing*, *307*, 72–77. https://doi.org/10.1016/j.neucom.2018.03.067

Copernicus. (2022). Index Corine Land Cover (clc). https://land.copernicus.eu/content/corine-land-cover-nomenclature-guidelines/html/

Daliakopoulos, I. N., Coulibaly, P., & Tsanis, I. K. (2005). Groundwater level forecasting using artificial neural networks. *Journal of Hydrology*, *309*(1-4), 229–240. https://doi.org/10.1016/j.jhydrol.2004.12.001

de Graaf, I., Gleeson, T., Rens van Beek, L., Sutanudjaja, E., & Bierkens, M. (2019). Environmental flow limits to global groundwater pumping. *Nature*, *574*, 90–94. https://doi.org/https://doi.org/10.1038/s41586-019-1594-4

DWD. (n.d.). Climate Data Center - Grids Germany- HYRAS dataset. https://opendata.dwd.de/climate_environment/CDC/grids_germany/daily/hyras_de/

Earman, S., & Dettinger, M. (2011). Potential impacts of climate change on groundwater resources - A global review. *Journal of Water and Climate Change*, *2*(4), 213–229. https://doi.org/10.2166/wcc.2011.034

Fabio, D. N., Abba, S. I., Pham, B. Q., Towfiqul Islam, A. R. M., Talukdar, S., & Francesco, G. (2022). Groundwater level forecasting in Northern Bangladesh using nonlinear autoregressive exogenous (NARX) and extreme learning machine (ELM) neural networks. *Arabian Journal of Geosciences*, *15*(7). https://doi.org/10.1007/s12517-022-09906-6

Fernando Nogueira. (2014). Bayesian Optimization: Open source constrained global optimization tool for Python.

Frick, C., Steiner, H., Mazurkiewicz, A., Riediger, U., Rauthe, M., Reich, T., & Gratzki, A. (2014). Central European high-resolution gridded daily data sets (HYRAS): Mean temperature and relative humidity. *Meteorologische Zeitschrift*, *23*(1), 15–32. https://doi.org/10.1127/0941-2948/2014/0560

Gholizadeh, H., Zhang, Y., Frame, J., Gu, X., & Green, C. T. (2023). Long short-term memory models to quantify long-term evolution of streamflow discharge and groundwater depth in Alabama. *Science of the Total Environment*, *901*. https://doi.org/10.1016/j.scitotenv.2023.165884

Goderniaux, P., Brouyère, S., Wildemeersch, S., Therrien, R., & Dassargues, A. (2015). Uncertainty of climate change impact on groundwater reserves - Application to a chalk aquifer. *Journal of Hydrology*, *528*, 108–121. https://doi.org/10.1016/j.jhydrol.2015.06.018

Gong, Y., Zhang, Y., Lan, S., & Wang, H. (2016). A Comparative Study of Artificial Neural Networks, Support Vector Machines and Adaptive Neuro Fuzzy Inference System for Forecasting Groundwater Levels near Lake Okeechobee, Florida. *Water Resources Management*, *30*(1), 375–391. https://doi.org/10.1007/s11269-015-1167-8

Guzman, S. M., Paz, J. O., & Tagert, M. L. M. (2017). The Use of NARX Neural Networks to Forecast Daily Groundwater Levels. *Water Resources Management*, *31*(5). https://doi.org/10.1007/s11269-017-1598-5

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Jordahl, K., Van den Bossche, J., Fleischmann, M., Wasserman, J., McBride, J., & Gerard, J. (2020). geopandas/geopandas: v0.8.1. https://doi.org/10.5281/zenodo.3946761

Kløve, B., Ala-Aho, P., Bertrand, G., Gurdak, J. J., Kupfersberger, H., Kværner, J., Muotka, T., Mykrä, H., Preda, E., Rossi, P., Uvo, C. B., Velasco, E., & Pulido-Velazquez, M. (2013). Climate change impacts on groundwater and dependent ecosystems. *Journal of Hydrology*, *518*(PB), 250–266. https://doi.org/10.1016/j.jhydrol.2013.06.037

LBEG. (2016). Hydrogeologische Räume und Teilräume in Niedersachsen. https://www.umwelt.niedersachsen.de/startseite/themen/wasser/grundwasser/grundwasserbericht_niedersachsen/nutzung_schutz_und_uberwachung/hydrogeologischer_uberblick/

LeCun, Y., Hinton, G., & Bengio, Y. (2015). Deep learning (2015), Y. LeCun, Y. Bengio and G. Hinton. *Nature*, *521*.

Lee, S., Lee, K. K., & Yoon, H. (2019). Using artificial neural network models for groundwater level forecasting and assessment of the relative impacts of influencing factors. *Hydrogeology Journal*, *27*(2), 567–579. https://doi.org/10.1007/s10040-018-1866-3

Lerner, D. N., & Harris, B. (2009). The relationship between land use and groundwater resources and quality. *Land Use Policy*, *26*(SUPPL. 1), 265–273. https://doi.org/10.1016/j.landusepol.2009.09.005

Linke, C. (2017). Leitlinien zur Interpretation regionaler Klimamodelldaten des Bund-Länder-Fachgespräches „Interpretation regionaler Klimamodelldaten“. http://www.lfu.brandenburg.de

Liu, Q., Gui, D., Zhang, L., Niu, J., Dai, H., Wei, G., & Hu, B. X. (2022). Simulation of regional groundwater levels in arid regions using interpretable machine learning models. *Science of the Total Environment*, *831*. https://doi.org/10.1016/j.scitotenv.2022.154902

LSN. (2016). *Öffentliche Wasserversorgung und Abwasserbeseitigung* (tech. rep.). https://www.statistik.niedersachsen.de/startseite/themen/umwelt_und_energie/umwelt-und-energie-in-niedersachsen-statistische-berichte-q-i-2-178924.html

Malik, A., & Bhagwat, A. (2021). Modelling groundwater level fluctuations in urban areas using artificial neural network. *Groundwater for Sustainable Development*, *12*(July 2020), 100484. https://doi.org/10.1016/j.gsd.2020.100484

Mohanty, S., Jha, M. K., & Raul, S. K. (2015). Using Artificial Neural Network Approach for Simultaneous Forecasting of Weekly Groundwater Levels at Multiple Sites, 5521–5532. https://doi.org/10.1007/s11269-015-1132-6

Moriasi, D. N., Gitau, M. W., Pai, N., & Daggupati, P. (2015). Hydrologic and water quality models: Performance measures and evaluation criteria. *Transactions of the ASABE*, *58*(6), 1763–1785. https://doi.org/10.13031/trans.58.10715

NMUEK. (2015). Lower Saxony contribution to the management plans 2015 to 2021 for the Elbe, Weser, Ems and Rhine river basins.

OSM. (2022). Download OpenStreetMap data for this region: Niedersachsen. https://download.geofabrik.de/europe/germany/niedersachsen.html

Pan, Y., Gong, H., Zhou, D., Li, X., & Nakagoshi, N. (2011). Impact of land use change on groundwater recharge in Guishui River Basin, China. *Chinese Geographical Science*, *21*(6), 734–743. https://doi.org/10.1007/s11769-011-0508-7

Pistocchi, A. (2015). *Leaf Area Index (MAPPE model)* (tech. rep.). European Commission, Joint Research Centre (JRC). https://data.jrc.ec.europa.eu/dataset/jrc-mappe-europe-setup-d-18-lai

Post, V. E., & von Asmuth, J. R. (2013). Revue : Mesure du niveau piézométrique-nouvelles technologies, pièges classiques. *Hydrogeology Journal*, *21*(4), 737–750. https://doi.org/10.1007/s10040-013-0969-0

Rajaveni, S. P., Brindha, K., & Elango, L. (2017). Geological and geomorphological controls on groundwater occurrence in a hard rock region. *Applied Water Science*, *7*(3), 1377–1389. https://doi.org/10.1007/s13201-015-0327-6

Rauthe, M., Steiner, H., Riediger, U., Mazurkiewicz, A., & Gratzki, A. (2013). A Central European precipitation climatology - Part I: Generation and validation of a high-resolution gridded daily data set (HYRAS). *Meteorologische Zeitschrift*, *22*(3), 235–256. https://doi.org/10.1127/0941-2948/2013/0436

Razafimaharo, C., Krähenmann, S., Höpp, S., Rauthe, M., & Deutschländer, T. (2020). New high-resolution gridded dataset of daily mean, minimum, and maximum temperature and relative humidity for Central Europe (HYRAS). https://doi.org/10.1007/s00704-020-03388-w/Published

Reback, J. e. a. (2020). pandas-dev/pandas: Pandas 1.4.2. https://zenodo.org/record/6702671#.YwEQwHZBzt8

Reichenau, T. G., Korres, W., Montzka, C., Fiener, P., Wilken, F., Stadler, A., Waldhoff, G., & Schneider, K. (2016). Spatial Heterogeneity of Leaf Area Index (LAI) and Its Temporal Course on Arable Land: Combining Field Measurements, Remote Sensing and Simulation in a Comprehensive Data Analysis Approach (CDAA). *PLoS ONE*, *11*(7), 158451. https://doi.org/10.1371/journal.pone.0158451

Retike, I., Bikše, J., Kalvāns, A., Dēliņa, A., Avotniece, Z., Zaadnoordijk, W. J., Jemeljanova, M., Popovs, K., Babre, A., Zelenkevičs, A., & Baikovs, A. (2022). Rescue of groundwater level time series: How to visually identify and treat errors. *Journal of Hydrology*, *605*. https://doi.org/10.1016/j.jhydrol.2021.127294

Roshni, T., Jha, M. K., & Drisya, J. (2020). Neural network modeling for groundwater-level forecasting in coastal aquifers. *Neural Computing and Applications*, *32*(16), 12737–12754. https://doi.org/10.1007/s00521-020-04722-z

Rust, W., Holman, I., Corstanje, R., Bloomfield, J., & Cuthbert, M. (2018). A conceptual model for climatic teleconnection signal control on groundwater variability in Europe. *Earth-Science Reviews*, *177*, 164–174. https://doi.org/10.1016/j.earscirev.2017.09.017

Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. *Advances in Neural Information Processing Systems*, *4*.

Takafuji, E. H. d. M., Rocha, M. M. d., & Manzione, R. L. (2019). Groundwater Level Prediction/Forecasting and Assessment of Uncertainty Using SGS and ARIMA Models: A Case Study in the Bauru Aquifer System (Brazil). *Natural Resources Research*, *28*(2), 487–503. https://doi.org/10.1007/s11053-018-9403-6

Tao, H., Hameed, M. M., Marhoon, H. A., Zounemat-Kermani, M., Heddam, S., Sungwon, K., Sulaiman, S. O., Tan, M. L., Sa'adi, Z., Mehr, A. D., Allawi, M. F., Abba, S. I., Zain, J. M., Falah, M. W., Jamei, M., Bokde, N. D., Bayatvarkeshi, M., Al-Mukhtar, M., Bhagat, S. K., . . . Yaseen, Z. M. (2022). Groundwater level prediction using machine learning models: A comprehensive review. *Neurocomputing*, *489*, 271–308. https://doi.org/10.1016/j.neucom.2022.03.014

Taylor, C. J., & Alley, W. M. (2001). Ground-water-level monitoring and the importance of long-term water-level data. *US Geological Survey Circular*, (1217), 1–68.

Treidel, H., Martin-Bordes, J. L., & Gurdak, J. J. (2011). *Climate change effects on groundwater resources: A global synthesis of findings and recommendations*.

UNESCO. (2020). The United Nations world water development report 2020: water and climate change.

UNESCO. (2022). *Water and climate change* (tech. rep.).

Van Der Walt, S., Colbert, S. C., & Varoquaux, G. (2011). The NumPy array: A structure for efficient numerical computation. *Computing in Science and Engineering*, *13*(2). https://doi.org/10.1109/MCSE.2011.37

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., . . . Vázquez-Baeza, Y. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wendt, D. E., Van Loon, A. F., Bloomfield, J. P., & Hannah, D. M. (2020). Asymmetric impact of groundwater use on groundwater droughts. *Hydrology and Earth System Sciences*, *24*(10). https://doi.org/10.5194/hess-24-4853-2020

Wunsch, A., Liesch, T., & Broda, S. (2018). Forecasting groundwater levels using nonlinear autoregressive networks with exogenous input (NARX). *Journal of Hydrology*, *567*, 743–758. https://doi.org/10.1016/j.jhydrol.2018.01.045

Wunsch, A., Liesch, T., & Broda, S. (2020). Groundwater Level Forecasting with Artificial Neural Networks: A Comparison of LSTM, CNN and NARX. *Hydrology and Earth System Sciences Discussions*, (March 2021), 1–23. https://doi.org/10.5194/hess-2020-552

Wunsch, A., Liesch, T., & Broda, S. (2021). Groundwater level forecasting with artificial neural networks: A comparison of long short-term memory (LSTM), convolutional neural networks (CNNs), and non-linear autoregressive networks with exogenous input (NARX). *Hydrology and Earth System Sciences*, *25*(3), 1671–1687. https://doi.org/10.5194/hess-25-1671-2021

Wunsch, A., Liesch, T., & Broda, S. (2022). Deep learning shows declining groundwater levels in Germany until 2100 due to climate change. *Nat Commun*, *13*. https://doi.org/https://doi.org/10.21203/rs.3.rs-420056/v1

Xu, Y. S., Shen, S. L., Cai, Z. Y., & Zhou, G. Y. (2008). The state of land subsidence and prediction approaches due to groundwater withdrawal in China. *Natural Hazards*, *45*(1), 123–135. https://doi.org/10.1007/s11069-007-9168-4

Yang, Y., & Hyndman, R. J. (2020). tsfeatures documentation. https://cran.r-project.org/web/packages/tsfeatures/vignettes/tsfeatures.html

Zanotti, C., Rotiroti, M., Sterlacchini, S., Cappellini, G., Fumagalli, L., Stefania, G. A., Nannucci, M. S., Leoni, B., & Bonomi, T. (2019). Choosing between linear and nonlinear models and avoiding overfitting for short and long term groundwater level forecasting in a linear system. *Journal of Hydrology*, *578*. https://doi.org/10.1016/j.jhydrol.2019.124015

Zhang, Q., Li, P., Ren, X., Ning, J., Li, J., Liu, C., Wang, Y., & Wang, G. (2023). A new real-time groundwater level forecasting strategy: Coupling hybrid data-driven models with remote sensing data. *Journal of Hydrology*, *625*. https://doi.org/10.1016/j.jhydrol.2023.129962

Table 2. Overview of geospatial features considered for the performance evaluation.

Feature	Description	Source
Distance to ~~water works~~ waterworks	Distance to water supply systems up to 10 km	OSM (2022)
Distance up to 25 km from the coastline	Distance to Lower Saxony coastline	OSM (2022)
CORINE land cover	~~Proportion in~~ Proportion in a 1 km ~~buffer~~ radius of the most relevant categories: (Non-irrigated arable land, pastures, coniferous forest, Discontinuous urban fabric)	Copernicus (2022)
Geomorphology	~~Proportion in 1km buffer~~ Proportion in 1 km radius of the most relevant categories: (Low Relief/medium-high SMI, sink areas/low-	BGR (2006)

Table 3. Overview of time series features considered for the performance evaluation.

Feature	**Description**	~~**Python Library**~~
Autocorrelation	~~Computed with a~~ Degree of similarity between a time series and a lagged version of itself. Here we used a lag of 6 ~~months.~~ time steps (6 months)	~~Partial autocor~~
~~Seasonal autocorrelation Autocorrelation without the seasonality.~~Stability	Variance of the means through over-lap-	~~Seasonal strength~~