Hydrol. Earth Syst. Sci., 29, 3405–3433, 2025
https://doi.org/10.5194/hess-29-3405-2025
© Author(s) 2025. This work is distributed under
the Creative Commons Attribution 4.0 License.

Hydrology and Earth System Sciences
Open Access

Towards a global spatial machine learning model for seasonal groundwater level predictions in Germany

Stefan Kunz[1], Alexander Schulz[2], Maria Wetzel[1], Maximilian Nölscher[1], Teodor Chiaburu[2], Felix Biessmann[2], and Stefan Broda[1]

[1]Federal Institute for Geosciences and Natural Resources, Berlin, Germany
[2]Berliner Hochschule für Technik, Berlin, Germany

Correspondence: Stefan Kunz (stefan.kunz@bgr.de)

Received: 7 November 2024 – Discussion started: 19 November 2024
Revised: 18 March 2025 – Accepted: 29 March 2025 – Published: 1 August 2025

Abstract. Reliable predictions of groundwater levels are crucial for sustainable groundwater resource management, which needs to balance diverse water needs and to address potential ecological consequences of groundwater depletion. Machine learning (ML) approaches for time series forecasting have shown promising accuracy for groundwater level prediction and, furthermore, offer scalability advantages over traditional numerical methods when sufficient data are available. Global ML architectures enable predictions across numerous monitoring wells concurrently using a single model, allowing predictions over a broad range of hydrogeological and meteorological conditions and simplifying model management. In this contribution, groundwater levels for 5288 monitoring wells across Germany were forecasted up to 12 weeks using two state-of-the-art ML approaches, the Temporal Fusion Transformer (TFT) and the Neural Hierarchical Forecasting for Time Series (N-HiTS) algorithm. The models were provided with historical groundwater levels, meteorological features, and a wide range of static features describing hydrogeological and soil properties at the monitoring sites. To determine the conditions under which the model achieves good performance and whether it aligns with hydrogeological system understanding, the model's performance was evaluated spatially and correlations with both static input features and time series features from hydrograph data were examined.

The N-HiTS model outperformed the TFT model, achieving a median Nash–Sutcliffe efficiency (NSE) of 0.5 for the 12-week prediction over all 5288 monitoring wells. Performance varied widely: 25 % of wells achieved an NSE > 0.68, while 15 % had an NSE < 0 with the best N-HiTS model. A tendency for better predictions in areas with high data density was observed. Moreover, the models achieved higher performance in lowland areas with distinct seasonal groundwater dynamics, in monitoring wells located in porous aquifers, and at sites with moderate permeabilities, which aligns with theoretical expectations. Overall, the findings highlight that global ML models can facilitate accurate seasonal groundwater predictions over large, hydrogeological diverse areas, potentially informing future groundwater management practices at a national scale.

1 Introduction

The growing availability of large datasets for hydrogeological applications has led to increased use of machine learning (ML) approaches in the hydrogeological domain for tasks such as groundwater level prediction (Tao et al., 2022; Collenteur et al., 2024) and groundwater level reconstruction (Chidepudi et al., 2024), as well as related tasks such as predicting groundwater recharge rates (Jung et al., 2024) or the 3D characterisation of aquifer systems (Manzoni et al., 2023). In recent years, an expanding field of research has demonstrated the potential of ML methods for groundwater level prediction, both in smaller areas (Guzman et al., 2017; Chidepudi et al., 2025; Wunsch et al., 2021, 2018) and for monitoring wells distributed throughout Germany (Heudorfer et al., 2024; Wunsch et al., 2022a). ML approaches aim to implicitly learn the underlying dynamics from measured data by modelling the statistical relationships between input features (e.g. precipitation) and target features (e.g. ground-

water levels). These models are based on the data at the respective groundwater monitoring wells and are thus easy to scale spatially and effective for groundwater level prediction over larger areas.

Single-well (local) models are the state of the art in ML-based groundwater level prediction, where for each monitoring well, an individual model has to be trained. In contrast, the class of so-called "global ML models" allows the creation of one ML model capable of predicting groundwater levels for multiple monitoring wells. These models train on different time series to learn underlying patterns shared across multiple time series, potentially enhancing the overall forecasting performance. Compared to single-well models, using multiple monitoring wells at once increases the diversity of training data, thereby widening the training envelope, which can result in improved performance during inference, e.g. because the model had access to rare observations such as extreme precipitation events (Kratzert et al., 2024). Furthermore, global ML approaches simplify model development and maintenance in comparison to single-well models. Besides incorporating time-dependent (dynamic) input features such as precipitation and temperature, these ML approaches can also make use of temporally static features that describe physical or environmental properties at the monitoring sites (e.g. permeability coefficient). These static features enable information sharing within groups of time series with similar static variable levels, aiming to enhance the model's ability to generalise to locations with similar static feature levels. By incorporating static features, these global models possess the potential for generalisation and regionalisation (Heudorfer et al., 2024; Kratzert et al., 2019). To assess the uncertainty in the predicted target feature from ML approaches for time series prediction, different methods are used, such as ensembling, Monte Carlo dropout (Althoff et al., 2021), quantile forecasting (Kan et al., 2022), or directly predicting a distribution of the target feature (Klotz et al., 2022), while in numerical groundwater models stochastic model calibration methods are often employed, which may involve sensitivity analyses of model parameters (Manzoni et al., 2024; Linde et al., 2017).

So far, in forecasting competitions, such as M4 (Makridakis et al., 2020) and M5 (Makridakis et al., 2022), global time series models have been successfully applied to forecasting problems in finance, retail, and economics and outperformed their competing local models. Recently, several new neural network architectures for global time series prediction were proposed. Among these architectures are DSSM (Rangapuram et al., 2018), DeepGLO (Sen et al., 2019), DeepAR (Salinas et al., 2020), StemGNN (Cao et al., 2021), TFT (Temporal Fusion Transformer) (Lim et al., 2021), N-HiTS (Neural Hierarchical Forecasting for Time Series) (Challu et al., 2022), and TiDE (Das et al., 2024). The recently developed TFT, a combination of recurrent neural networks and self-attention layers, and N-HiTS, a time series decomposition algorithm based on multilayer perceptrons (MLPs), have shown promising predictive capabilities. The TFT architecture has already been successfully applied in the environmental sciences, e.g. for forecasting CO_2 emissions from 1 h up to 1 week using about 6 million CO_2 measurements and other time-dependent (dynamic) variables (Linardatos et al., 2023). Furthermore, the TFT was also used for streamflow prediction on 2610 basins across the world using 38 dynamic and 211 static features, thereby outperforming models based on long-short term memory (LSTM) and other transformer architectures (Koya and Roy, 2023). In environmental research, the N-HiTS architecture has been successfully applied to predict wastewater levels in rain tanks over a 10 h prediction horizon (Chiaburu and Bießmann, 2024). Furthermore, N-HiTS was compared to other models in a study on global weather forecasting (Wu et al., 2023).

For predicting groundwater levels in Germany, a recent study by Heudorfer et al. (2024) used a global LSTM architecture integrating both dynamic (meteorological) and static input features. Their optimal models achieved a Nash–Sutcliffe efficiency (NSE) of above 0.8 in the test period across a selected set of 108 monitoring wells, which were relatively evenly distributed throughout Germany. This study seeks to expand upon this research by using a larger number of monitoring wells and exploring more complex ML architectures. The key contributions of this study include (1) introducing and evaluating modern ML architectures for groundwater modelling, complementing classical architectures such as LSTM; (2) providing a comprehensive empirical evaluation of various models performance using a dataset larger than those of previous studies; (3) investigating the importance of input features, with a specific focus on static features, in influencing the prediction accuracy; and (4) assessing the capability of these models to align with hydrogeological system understanding. We hypothesise that the large amount of time series data used in combination with static input features will help to improve predictive performance and enhance hydrogeological system understanding. In this study, the two state-of-the-art global ML time series architectures, TFT and N-HiTS, are applied to seasonally predict groundwater levels up to 12 weeks for the so far most comprehensive groundwater dataset covering large parts of Germany (5288 monitoring wells). To the best of our knowledge, this is the first study that attempts to make groundwater level predictions on a national scale for such a high number of monitoring wells and one of the first studies that uses global ML architectures to predict groundwater levels.

The remainder of the paper is organised as follows: first we introduce the dataset (Sect. 2.1) and the deep learning architectures (Sect. 2.2) along with the experimental design (Sect. 2.3). Then, the predictive performance is reported for both architectures and analysed with respect to factors affecting the predictive performance. To quantify the impact of static features on the prediction accuracy, model variants with and without static features are examined for both architectures. Moreover, the interpretable nature of the TFT is

used to understand the importance of the input features and important past time steps for the model performance. Finally, the results are discussed in terms of general and spatial model performance as well as the feature importance to demonstrate the potential of modern ML approaches for groundwater level prediction and to determine if the models reflect our understanding of the hydrogeological system (Sects. 3 and 4).

2 Data and methods

2.1 Data

2.1.1 Groundwater level time series and preprocessing

Groundwater level measurements in the period from 1990 to 2016 were provided by the environmental agencies and the geological surveys of most of the German federal states at a weekly resolution. The preprocessing of the groundwater levels was kept simple, mainly for two reasons: first, simple and few preprocessing steps allow retaining as many training data points as possible; second, it allows assessing the predictive performance under realistic conditions with heterogeneous data distributions. For each groundwater level time series jumps between two time points greater than 50 times the average change of the given time series were removed and gaps up to 4 weeks were linearly interpolated. Time series with gaps greater than 4 weeks were truncated at the gap.

The preprocessing resulted in groundwater time series of 5288 monitoring wells, which corresponds to approximately 4.5 million records for model training. These monitoring wells are distributed throughout Germany, with the exception of some federal states, where only few (Mecklenburg–Western Pomerania, Bremen, Hamburg) or no monitoring wells (Thuringia and Saarland) satisfied the preprocessing criteria. Nevertheless, all major hydrogeological and climatic conditions of Germany are represented in the dataset. For example, in the hydrogeological district encompassing Thuringia, common aquifer types include fractured, fractured and karstified, and fractured and porous aquifers. Monitoring wells located in these aquifer types are also present in other regions covered in the dataset (e.g. over 612 monitoring wells are located in fractured aquifers; Table B7).

2.1.2 Dynamic features

An overview of the dynamic input features can be found in Table 1. Besides being the target to predict, the groundwater level was also used as a dynamic input feature. Further dynamic input features were precipitation, relative humidity, and temperature, which have been successfully used in previous modelling studies (see Heudorfer et al., 2024, and citations therein) and should theoretically exert a strong influence on groundwater levels. These meteorological dynamic features were extracted from HYRAS 5.0 (Razafimaharo et al., 2020). Moreover, a sinusoidal curve fitted to the temperature was used, which was a good predictor for groundwater levels in previous studies (Wunsch et al., 2021). The HYRAS dataset is a raster product published by the National German Weather Service and holds gridded meteorological data for Germany over the period from 1951 to 2022 for temperature, to 2023 for humidity, and to 2024 for precipitation (status as of October 2024). Additional dynamic input features were the leaf area index (LAI) (Pistocchi, 2015) and the day of the year expressed in sine and cosine values.

The meteorological input features and the LAI were extracted within a 1 km buffer around the groundwater wells. Thereby, a weighted average was calculated based on the area covered by the pixels within the buffer.

2.1.3 Static features

The static features used in the study cover environmental characteristics from the domains hydrogeology, soil, topography, and land cover (see Table 2). They describe dominant physical factors that control groundwater dynamics at the well locations, aiming to facilitate the generalisation of relationships between monitoring sites with similar environmental characteristics. All static input features except the landform Shannon index and the elevation were extracted within a 1 km buffer around the groundwater wells. For the values of the numerical features, a weighted average was calculated based on the area covered by the pixels within the buffer. For categorical features, the category with the largest area share within the buffer was used. The landform Shannon index and elevation were extracted directly at the well. The topographic wetness index (TWI) was calculated according to Beven and Kirkby (1979). Additionally, the mean and the standard deviation of each groundwater level time series computed on the respective training data periods were used as static input time series features.

2.2 Global machine learning algorithms

2.2.1 Temporal Fusion Transformer (TFT)

The TFT is an attention-based neural network architecture for time series forecasting and has achieved good results on various forecasting tasks, often outperforming other models like LSTMs and transformers (Lim et al., 2021). The main building blocks of the TFT are gating mechanisms, variable selection networks (VSNs), static covariate encoders, LSTM encoder–decoders, and multi-head self-attention layers.

The gating mechanisms regulate the degree of non-linear processing of the input by employing gated residual networks (GRNs), allowing the model to skip entire layers when necessary. VSNs help to remove noisy inputs and can be used to assess the importance of input features. Thereby, feature importance is assessed separately for static and dynamic inputs (i.e. past and future inputs) on the test set. At each time step, a

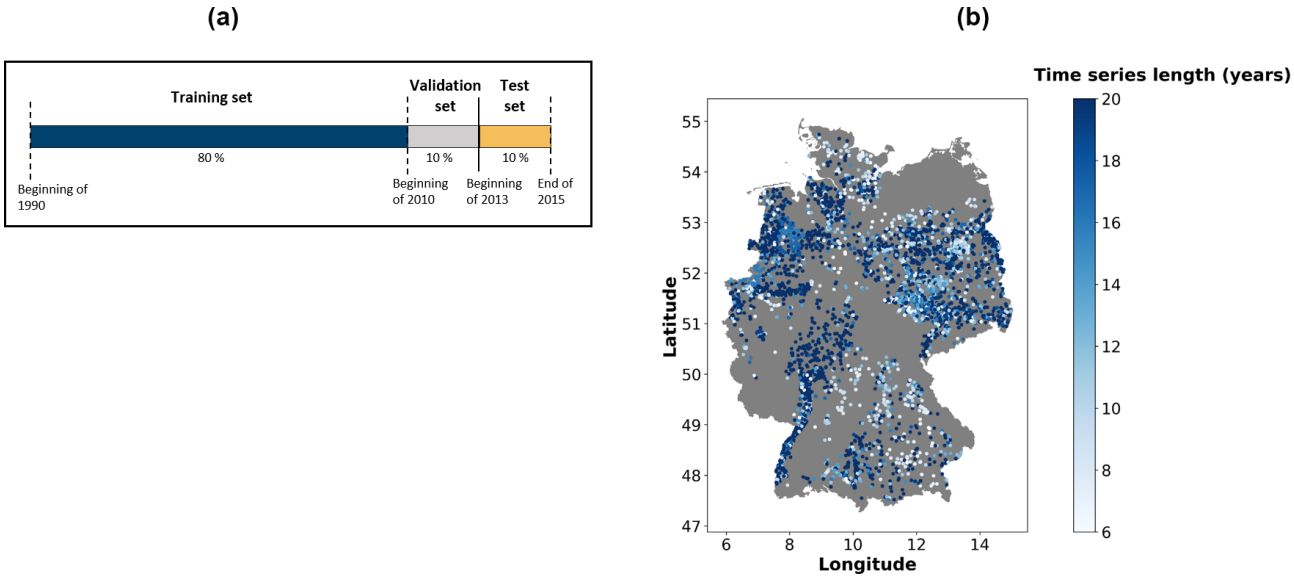

Figure 1. **(a)** Split into training, validation, and test data. **(b)** Spatial distribution of the 5288 wells used for groundwater level prediction. The colour scale indicates the length of the individual groundwater level time series in the training dataset.

Table 1. Description of the dynamic input features.

Type	Name	Description	Source
Hydrogeology	Groundwater levels	Groundwater hydrographs aggregated to weekly values	Provided by the environmental agencies and geological surveys in Germany (unpublished)
Meteorology	Temperature	Mean temperature (°C), daily values aggregated to weekly, grid size: 5 km × 5 km	Razafimaharo et al. (2020)
	Sinus temperature	–	
	Precipitation	Sum of precipitation (mm), daily values aggregated to weekly, grid size: 1 × 1 km	Razafimaharo et al. (2020)
	Relative humidity	Mean relative humidity (%), daily values aggregated to weekly, grid size: 5 × 5 km	Razafimaharo et al. (2020)
Vegetation	Leaf area index (LAI)	The LAI describes the area of leaves per unit ground area and is related to the evapotranspiration in an area. It is reported in monthly averaged values. Grid size: 1 × 1 km	Pistocchi (2015)
Annual cycle	Sinus day	–	
	Cosinus day		

flattened feature vector is fed into a GRN followed by a Softmax layer, producing an output vector with so-called variable selection weights. The importance of each feature is then represented by the average of the variable selection weights over all time steps. After the VSN is applied, static covariate encoders integrate static information with the dynamic information using a GRN. The dynamic outputs of the VSN are passed to the LSTM encoder–decoder, which captures short-term dependencies by learning past data representations (encoder) to predict future values (decoder). Self-attention is then applied to the output by the LSTM encoder–decoder to weigh the different time steps of input sequences and dynamically adjust how they affect the output. Self-attention layers enable the model to identify important time steps and approximate long-term dependencies across the input sequences. The attention weights can be visualised for interpretability.

Table 2. Description of the static input features.

Type	Name	Description	Source
Hydrogeology	Hydrogeological spatial structure of Germany	Classification of areas with similar hydrogeological characteristics, groundwater conditions, and geologic genesis in Germany; the 10 major hydrogeological districts are used	BGR and SGD (2019)
	Aquifer type	Five aquifer type categories (e.g. porous or karstified)	BGR and SGD (2019)
	Permeability coefficient	Six categories (e.g. high ($>$ 0.001–0.01 $m\,s^{-1}$))	BGR and SGD (2019)
	Groundwater recharge	Mean annual groundwater recharge rates 1961–1990 ($mm\,a^{-1}$)	BGR (2019)
	Mean groundwater level	–	
	Standard deviation groundwater level	–	
Soil	Soil texture	13 categories (e.g. sandy loam)	BGR (2007)
Land cover	Land cover	Nine categories (e.g. herbaceous vegetation)	Buchhorn et al. (2020)
Topography	Topographic wetness index (TWI)	Terrain-driven wetness potential	Beven and Kirkby (1979)
	Elevation	Elevation at the groundwater well	European Environment Agency (2018)
	Shannon index landforms	Diversity index of landform types: higher values indicate more landform types and/or landform types having more similar proportions within the aggregation window (10×10 km)	Amatulli et al. (2018)
	Divide to stream distance (EUMOHP_DSD1)	Distance from hypothetical groundwater catchment divide to nearest stream (hydrologic order 1) at the monitoring well location	Nölscher et al. (2022)
	Lateral position (EUMOHP_LP1)	Relative position of the monitoring well laterally along the divide-to-stream stretch (hydrologic order 1)	Nölscher et al. (2022)
	Stream distance (EUMOHP_DS1)	Distance from the monitoring well to the nearest stream (hydrologic order 1)	Nölscher et al. (2022)

2.2.2 Neural Hierarchical Interpolation for Time Series Forecasting (N-HiTS)

N-HiTS (Challu et al., 2022) is a state-of-the-art deep learning model for time series forecasting, improving upon its predecessors, the N-BEATS and N-BEATSx models (Oreshkin et al., 2020; Olivares et al., 2022), by enhancing long-term forecasting capabilities and computational efficiency. N-HiTS outperformed its predecessors and several transformer-based methods on various forecasting tasks, especially for long-term horizons, and also required less memory and computational power (Challu et al., 2022).

The N-HiTS architecture is composed of so-called blocks organised into different stacks. Each block consists of a MaxPool layer and several multilayer perceptrons (MLPs). The MaxPool layers subsample the series at different resolutions (e.g. daily, weekly), allowing each stack to focus on different frequencies (short-term to long-term) within the data. Frequencies are projected onto a basis of simple functions such as sine waves and step functions (related concepts are short-time Fourier transform or wavelets). Each subsampled version of the series is processed by a series of MLPs that learn patterns within the data at their respective frequencies. Each MLP block outputs a backcast and forecast, where the backcast is subtracted from the input and the remaining signal is passed to the next block through residual connections. The partial predictions obtained from each block are combined through hierarchical interpolation. The final predictions consist of the summed-up partial predictions from each stack.

2.3 Experimental design and training

For every well, groundwater level time series were split into a training, validation, and test dataset. The validation and test

datasets comprise a period of 3 years starting in 2010 and 2013, respectively. The training data only included time series that were at least 6 years long and available from 2004 onwards to the validation period starting in 2010 (Fig. 1). Thus, the length of the training period varied (descriptive statistics in Table A1), whereby 54 % of all time series spanned the entire period of available data until 1990 (Fig. 1). Time series characteristics such as the mean and standard deviation of all included groundwater level time series were relatively comparable between the training, validation, and test period (Fig. A1). The exogenous dynamic features were divided into training, validation, and test data based on the split of the groundwater level time series.

In order to assess the impact of the static features on the predictive quality of the global models, two model variants were considered: the models were provided with (1) all dynamic and static feature and (2) solely dynamic features (referred to as purely dynamic models hereafter). For comparison, TFT models using exogenous features but excluding historical groundwater levels as input features were trained for both variants. However, the models' performance was much worse compared to the models trained with the groundwater levels as an input feature. Hence, we decided not to pursue further analyses in this regard. The results for this analysis are reported in Appendix B (Table B6). As a baseline for the 1-week prediction horizon, a naïve forecast was used.

Both ML architectures used in this study are designed for sequence-to-sequence predictions. During training, the models processed an input sequence autoregressively and predicted an output sequence of groundwater levels. For each time step, a look-back window (i.e. sequence length) of 52 weeks for the dynamic features was used to represent one annual cycle. Groundwater levels were predicted for 12 weeks. During the 12-week prediction the model has access to the exogenous dynamic features, but not to the groundwater level (Fig. 2).

All model variants were trained for 10 epochs and with 10 different random seeds to account for the stochasticity in the initialisation of model weights. Large batch sizes were used (TFT: 4096, N-HiTS: 1024) to avoid overfitting and to accelerate the training. The risk of overfitting was further reduced by the application of early stopping on the validation loss, a dropout rate of 0.2, and learning rate scheduling using stochastic weight averaging after the second epoch (Izmailov et al., 2019). Thus, the selected hyperparameter values were based on empirical heuristics recommended by domain experts, aiming to reduce overfitting and minimise training time. All model variants were trained for a maximum of 10 epochs, a duration sufficient to ensure model convergence. In many cases, training terminated earlier due to the implementation of the early stopping criteria. During training, models were optimised with the Ranger optimiser (Wright and Demeure, 2021). For both architectures the multivariate quantile loss function was used (Kan et al., 2022). Thereby, each model learns to predict different parts

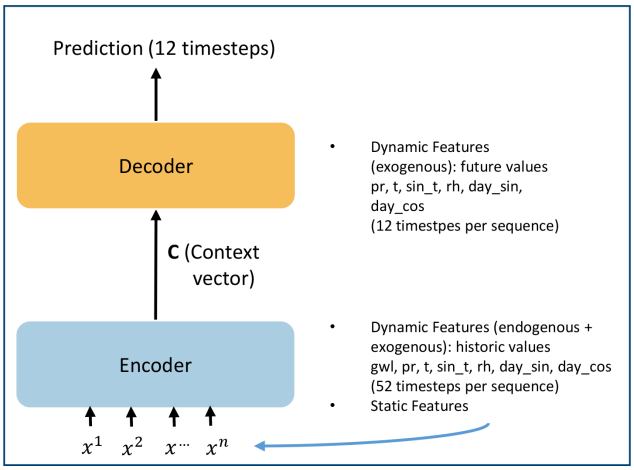

Figure 2. Abstract depiction of the modelling process for seasonal groundwater level predictions. Both ML architectures used follow an encoder–decoder structure. The encoder creates a latent representation of the input features, which is denoted as a context vector ($\boldsymbol{c}$). The decoder is conditioned on this context vector to generate predictions. Historical values of groundwater levels (gwl), precipitation (pr), temperature (t), sinus of the temperature (sin_t), relative humidity (RH), the annual cycle (day_sin, day_cos), and the static features are processed as sequences in the encoder (input per time step denoted by x). In the decoder, the exogenous dynamic features are used for the prediction horizon, while groundwater levels are not used.

of the conditional distribution simultaneously. Accordingly, prediction intervals were obtained with one model. The quantile loss was chosen because it has a linear relationship with the error magnitude, making it more robust towards outliers than metrics with a quadratic relationship, such as the mean squared error (MSE). Outliers may arise e.g. from extreme precipitation events.

2.4 Model evaluation

For the predicted groundwater levels, the 0.5 quantile is reported as a point forecast. The prediction intervals are based on the 0.10 and 0.90 quantiles. All models were evaluated with the NSE on the test data. In addition, the root mean square error (RMSE), the relative mean bias error (rMBE), and, for evaluating prediction intervals, the interval score are reported for comparison. The rMBE is defined as the mean deviance divided by a constant, for which the standard deviation of the groundwater levels is chosen. The interval score is a proper scoring rule that combines two properties of a prediction interval: (1) the sharpness, which refers to the concentration of the predictive distribution and is a property of the prediction only, and (2) the calibration, which denotes how much the true value is outside of the predicted interval (Gneiting and Raftery, 2007). Smaller values indicate a better estimate and a narrower prediction interval. Details on the calculation of both metrics are given in Appendix Sect. A3.

To compare model performances, we aggregated the error metrics to the median for each well across the 10 initialisations. Additionally, we also show the median NSE values across the 5288 monitoring wells for each individual initialisation (Fig. 4).

To investigate if certain hydrogeological conditions allow better predictions, the NSE values of each monitoring well were evaluated for a prediction horizon of 12 weeks across the categorical static features, including the large hydrogeological districts of Germany (Fig. B5). Additionally, Spearman correlations of the static numeric features and the NSE of each monitoring well were calculated. Areas with a high data density were identified using kernel density estimation (KDE), and median NSE values for the monitoring wells within these areas were compared.

To examine how the ML approaches performed with regard to groundwater hydrographs exhibiting differing behaviours, NSE values were correlated with a number of time series features describing the groundwater dynamics at each monitoring well. The time series features used were groundwater hydrograph specific features derived by Wunsch et al. (2022b), namely the seasonal behaviour (agreement with the expected seasonality, i.e. max in March and min in September), the flashiness (frequency of short-term changes), and the range ratio (superimposed long-period signals). The calculation of these features followed Wunsch et al. (2022b). Furthermore, standard statistical time series features were employed, namely the amplitude (range), skewness, linear trend (slope), and length of the time series (including the test period). For the linear trend, a linear regression was fitted and the obtained slope was correlated with the NSE, whereby non-significant slopes were set to zero. The nonparametric Spearman correlation coefficient was used, given that the NSE and the features considered for correlation analysis did not seem to satisfy the bivariate normality assumption of the Pearson correlation coefficient.

The TFT architecture offers interpretable insights via the variable selection weights and attention scores. For the predictive performance of the TFT, the feature importance ranking is reported as well as important past time steps.

3 Results

3.1 General model performance

Both global ML architectures achieved a high performance for the 1-week prediction horizon with a median NSE of 0.91 (TFT) and 0.92 (N-HiTS) for all 5288 monitoring wells. The naïve forecast (baseline) for the 1-week prediction achieved a median NSE of 0.91 (Table B3). ML architectures outperformed the baseline by at least 0.1 NSE in 8 %–10 % of the monitoring wells, depending on the ML architecture (Table B4). These wells exhibited greater variability in their groundwater level dynamics, indicated by higher flashiness values compared to the overall dataset. In contrast, the naïve forecast surpassed the ML architecture performances by at least 0.1 NSE in 0.07 %–0.7 % of monitoring wells (Table B5). For the 12-week prediction horizon, the N-HiTS models consistently outperformed the TFT models based on the NSE. The N-HiTS model provided with static and dynamic features achieved a median NSE of 0.5 across the 5288 monitoring wells (TFT 0.34, Fig. 3, Table B1), meaning that for approximately 81 % (4286) of the monitoring wells N-HiTS achieved higher NSE values than the TFT. A wide spread in the model performance was observed for all model variants; e.g. the N-HiTS model with static and dynamic input features achieved for 25 % of the monitoring wells an NSE of 0.68 or greater (TFT 0.55 or greater). On the other hand, an NSE below zero was achieved for about 15 % of the monitoring wells, indicating that the model predictions were a worse approximation to the observed groundwater levels than the mean of the observed values (TFT 22 % of the monitoring wells). Despite relatively similar NSE values at the 1-week prediction horizon, the NSE decreased more prominently for the TFT model variants than for the N-HiTS model variants over the longer prediction horizons. Generally, model performances decreased approximately linearly across the prediction horizons, and the patterns observed for the 1- and 12-week prediction horizons are representative for the other prediction horizons (Fig. 4). Moreover, the performance of the TFT models was more influenced by the initialisation of the model weights compared to the N-HiTS models, indicated by the larger spread of median NSE values per initialisation (Fig. 4).

In terms of the RMSE and rMBE the N-HiTS models performed slightly better than the TFT models (Fig. 3). For the 12-week prediction horizon, N-HiTS provided with static and dynamic features achieved a median rMBE of 0.13 m (TFT 0.15 m) and median rMBE of 0.08 m (TFT 0.13 m). For the RMSE and rMBE, the differences between the purely dynamic model variants and the models provided with all dynamic and static input features were mostly small at around 0.01 m.

Regarding the interval score (Fig. 3), the N-HiTS models achieved lower values than the TFT models, indicating that smaller and more precise prediction intervals could be obtained with N-HiTS. For the 12-week prediction horizon, the purely dynamic model variants had smaller interval scores than the model variants provided with static and dynamic input features (Table B2). N-HiTS (purely dynamic) achieved an interval score of 3.68 (TFT (purely dynamic) 4.41), while the interval score for the N-HiTS model provided with static and dynamic features was 3.89.

The addition of static features marginally improved the performance of the N-HiTS architecture (the difference in median NSE was 0.03 from the purely dynamic model for the 12-week prediction, $P < 0.001$) and did not improve the performance of the TFT architecture (Fig. 3, Table B1).

3.2 Spatial model performance

Investigating how the prediction accuracy is distributed among the groundwater monitoring wells reveals spatial patterns and variations across Germany. Both TFT and N-HiTS models showed similar spatial patterns in the predictive performance (Fig. B2). For the sake of simplicity, we focus on the results of the best-performing model, the N-HiTS model provided with static and dynamic features. As mentioned earlier, for the 1-week prediction horizon there were almost exclusively forecasts with a high prediction accuracy. However, poor performances with a median NSE below zero were observed for wells primarily located in southeastern Germany. For the 12-week horizon, there are regions where the model forecasts appear to be generally better and above the median NSE of 0.5, particularly in the Upper Rhine Graben, northwestern Germany, and large parts of northeastern Germany (Figs. 5, B2). Notably, better predictions were often achieved in areas with higher data density, though exceptions exist. The KDE analysis identified eight areas with high data density (Fig. 6). Among them, five show a better performance than the median NSE of 0.5 (median NSE between 0.59 and 0.78). However, three high-data-density areas, located in the southeast of the North and Central German Unconsolidated Rock District, central south of the North and Central German Unconsolidated Rock District, and the west of the Alpine Foreland, had median NSE values between 0.29 and 0.32. The Spearman correlation between the NSE and the KDE values was 0.12. Overall, there was no clear visual spatial distribution of NSE values; i.e. monitoring wells with poor model prediction accuracy (NSE below 0) were distributed across Germany and also occurred in close proximity to sites with better prediction accuracy, resulting in an uneven spatial pattern of model performance.

3.3 Model performance comparison across static features and time series features

To understand under which conditions the ML models perform best and to assess whether they capture influential factors affecting groundwater level dynamics, an analysis of the Spearman correlation between model performance and both numeric static features and time series features of the groundwater hydrographs was carried out. The correlations were largely similar for the N-HiTS and TFT model variants; hence results for the N-HiTS model provided with static and dynamic features for the 12-week prediction are reported in the text, if not denoted differently (Table 3). Similarly, regarding the performance distribution across the different static feature categories, the results for the best-performing model, N-HiTS provided with static and dynamic features, are reported.

No strong correlations between the static numeric features and the NSE values for the 12-week prediction were found. The most notable correlations were a positive correlation between the TWI and the NSE ($\rho = 0.22$) and negative correlations between elevation ($\rho = -0.27$) and the landform Shannon index ($\rho = -0.23$) with the NSE. For the time series features, the highest correlations were obtained for the seasonal behaviour (Table 3). Here, a positive correlation of $\rho = 0.28$ was obtained ($\rho = 0.34$ for N-HiTS, purely dynamic). Other notable correlations were a negative correlation between NSE and the flashiness of the groundwater hydrographs ($\rho = -0.21$) and a positive correlation with the range ratio ($\rho = 0.17$). A weak correlation was found between the NSE values and the linear trend as well as the NSE and the time series length ($\rho <= 0.03$). All reported correlations were statistically significantly different from zero (Table 3). Distributions of the time series features are provided in Fig. B3. A wide variation in model performance was observed across the different static feature categories (Fig. B4, Table B7). The highest median NSE values were achieved for monitoring wells situated in porous aquifers (0.52), at sites with moderate to high permeability coefficients of $> 1 \times 10^{-5}$ to $1 \times 10^{-3}\,\mathrm{m\,s^{-1}}$ (0.56), within closed deciduous broadleaf forests (0.60), and in peatlands (0.66). A total of 83 % of the wells were located in porous aquifers, and 68 % were positioned in areas with moderate to high permeability. Only a small proportion of wells were located within deciduous forests (4 %), primarily in the Upper Rhine Graben, or in peatlands (5 %), mostly in northern Germany (Figure B6). For the different major hydrogeological districts of Germany, the best results for the 12-week predictions were obtained in the North and Central German Unconsolidated Rock District, the Upper Rhine Graben, and the Southwest German Basement (median NSE between 0.56 and 0.7). In these districts many wells were located, they had a high density of wells, or they were located near districts with a high density of wells.

3.4 Feature importance and attention of the TFT models

The feature importance values of the TFT models derived from the variable selection networks were dependent on the model weight initialisations and showed a relatively high spread for most of the input features (Fig. 7). The feature importance ranking was dominated for the dynamic input variables by the groundwater level. Other important dynamic features were precipitation, temperature, and humidity, though for some initialisations the day of the year was also important (Figs. 7, B9). Among the static inputs, the standard deviation of the groundwater level was the most important input feature, followed by the TWI, the aquifer type, and the landform Shannon index.

The most important past time steps, according to the attention scores, were often at the beginning of the input sequences (52 weeks, i.e. the week a year ago) and recent time points. For the 1-week prediction this was often the last time point and for the 12-week prediction this was often 12 weeks before the prediction (Figs.B10, B11).

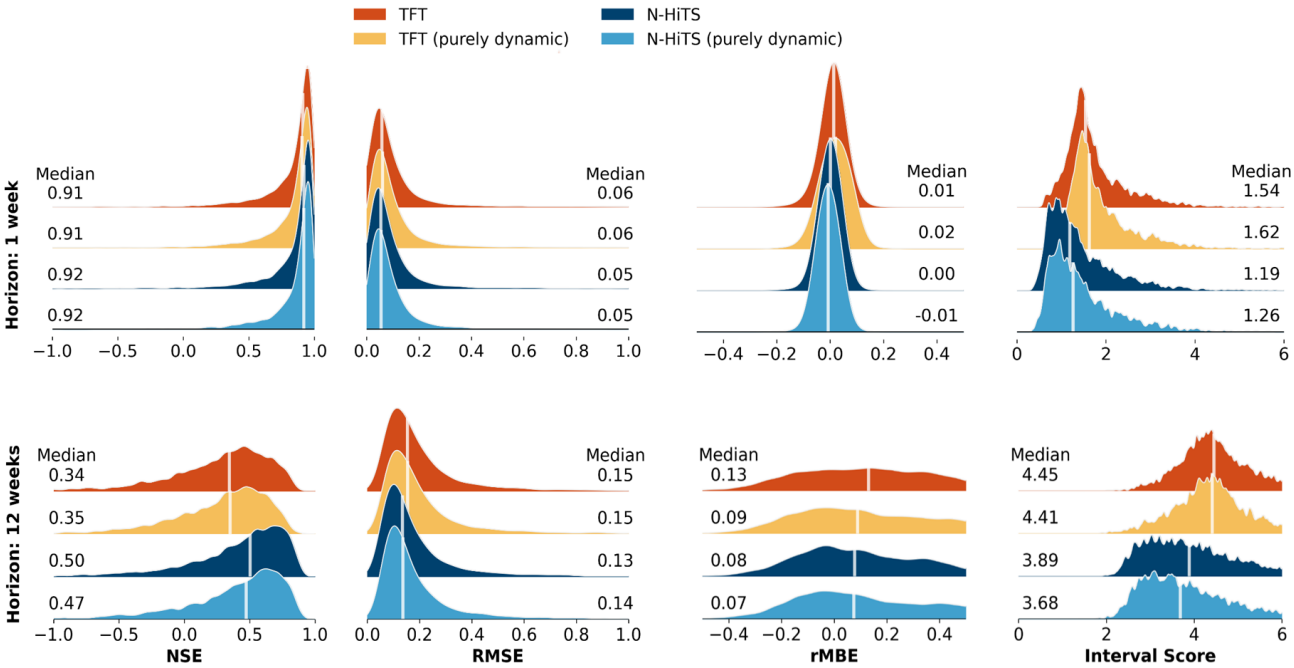

Figure 3. Overview of the predictive performance of the global models for the model variants with all features and the purely dynamic model variants. Distributions are shown of NSE, RMSE, rMBE, and interval score values achieved for the 5288 monitoring wells. Reported values are based on the median for each metric of the 10 initialisations for each well. White lines denote the median value of each distribution.

Table 3. Spearman correlations between the NSE values for the 12-week prediction horizon per monitoring well and the numeric static input features as well as the time series features. Correlations between NSE values for the purely dynamic model variants and the static numeric features were omitted because these features were not an input to these models. Significance codes: three asterisks (***) indicate 0, two asterisks (**) indicate <= 0.01, one asterisk (*) indicates <= 0.05, and no asterisk indicates > 0.05.

ρ_{Spearman}	NSE TFT	NSE TFT (purely dynamic)	NSE N-HiTS	NSE N-HiTS (purely dynamic)
Static numeric feature				
Elevation	−0.24***		−0.27***	
EUMOHP_DSD1	0.09**		0.1**	
EUMOHP_LP1	0.1**		0.1**	
EUMOHP_SD1	0.14**		0.14**	
Groundwater recharge	0.14**		0.1**	
Shannon index landforms	−0.23***		−0.23***	
TWI	0.18***		0.22***	
Time series feature				
Seasonal behaviour	0.27**	0.30**	0.28**	0.34**
SDdiff (flashiness)	−0.22***	−0.23***	−0.21***	−0.19***
Range ratio	0.1**	0.09**	0.17***	0.21***
Amplitude	−0.12**	−0.1**	−0.13**	−0.12**
Skew	−0.09**	−0.07**	−0.11**	−0.11**
Linear trend (slope)	0.009*	0.03**	−0.02**	−0.02**
Time series length	0.01*	−0.002	0.02**	0.03**

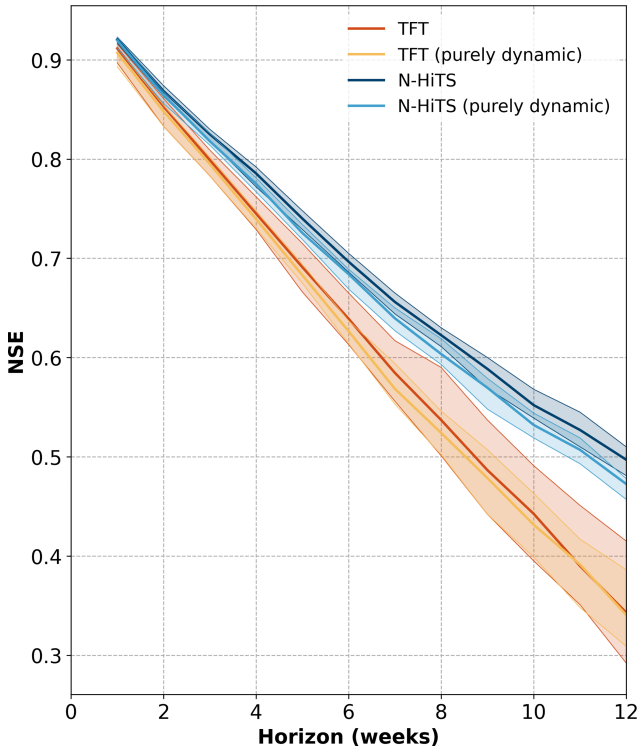

Figure 4. Range of median NSE values of the 10 initialisations (shaded colours) for each model variant for the 12-week prediction horizons. The bold lines represent the median of the 10 initialisations.

4 Discussion

4.1 General and spatial model performance

The results for the 12-week prediction demonstrate that the tested global ML approaches can provide good predictions for a large number of monitoring wells distributed over a broad region with diverse hydrogeological properties. Thereby, the N-HiTS architecture outperformed the TFT architecture by achieving a median NSE of 0.5 across all 5288 monitoring wells. Both models show high prediction accuracies for a considerable number of wells when an NSE of at least 0.7 is considered to be indicative of high performance (see Wunsch et al., 2022a). Precisely, the N-HiTS model provided with static and dynamic features achieved an NSE of at least 0.7 for 1163 monitoring wells (approximately 22 % of all monitoring wells) and the TFT model achieved an NSE of at least 0.7 for 489 monitoring wells (9 % of all monitoring wells). Overall, the N-HiTS model showed a good level of model performance, given the high number of monitoring wells included without preselection and the heterogeneous hydrogeological situation of the study area, comprising 10 different major hydrogeological districts and 35 hydrogeological regions (BGR and SGD, 2019).

Two previous studies have predicted groundwater levels for up to 118 monitoring wells distributed across Germany using single-well models based on convolutional neural networks (CNNs) (Wunsch et al., 2022a) and a global LSTM architecture (Heudorfer et al., 2024). A total of 82 of these 118 monitoring wells were also used in our study. Nevertheless, direct comparisons to our study should be made cautiously due to differences in input features and experimental design. The single-well models solely used meteorological input features (Temperature and precipitation), while the LSTM approach included static features; however, in both studies, groundwater level was not used as an input feature. The wells in these studies were preselected on the basis that their groundwater dynamics were primarily influenced by climatic processes and could be accurately predicted in the past. In addition, the lengths of the training, validation, and test period were different compared to our study, and Heudorfer et al. (2024) also used partly different static input features. The single-well models achieved an NSE of at least 0.7 (median NSE = 0.81) and the global LSTM approach a median NSE of 0.82 for the 1-week prediction horizon. For 82 of the 118 monitoring wells, our best model, the N-HiTS model provided with static and dynamic features, achieved an NSE of 0.93 for the 1-week prediction horizon and an NSE of 0.62 for the 12-week prediction horizon. Thus, the N-HiTS model can compete with the other approaches and is suitable for the purpose of short-term groundwater level prediction. However, N-HiTS in its current implementation requires the target feature as input feature and is for this task inferior to the single-well CNNs or the global LSTM.

The N-HiTS model produced narrower prediction intervals based on the interval score, with the purely dynamic model achieving narrower prediction intervals than the model provided with static and dynamic features for the 12-week prediction horizon. Example hydrographs included in Appendix B illustrate these results (Fig. B1). This finding implies that when the models have access to static features, they generalise or cluster across similar static feature levels, encountering a wide range of groundwater dynamics. Consequently, prediction intervals may be wider than in purely dynamic models, which estimate the 0.1 and 0.9 quantiles only based on historical groundwater levels and climatic data. This result indicates that the static features currently lack sufficient "distinguishing potential" and may generalise across wells that are actually environmentally distinct. One possible reason for this could be uncertainties in the static input features, which are discussed further in Sect. 4.3. Moreover, adding other relevant features, such as lithostratigraphic data at the monitoring well (which were unavailable for this study), could mitigate this issue.

The results further suggest that areas with a higher data density tend to have a higher prediction accuracy. Poor model performances in high-data-density regions may be explained by specific regional and anthropogenic factors, e.g. former and recent activities of lignite mining in the southeast of

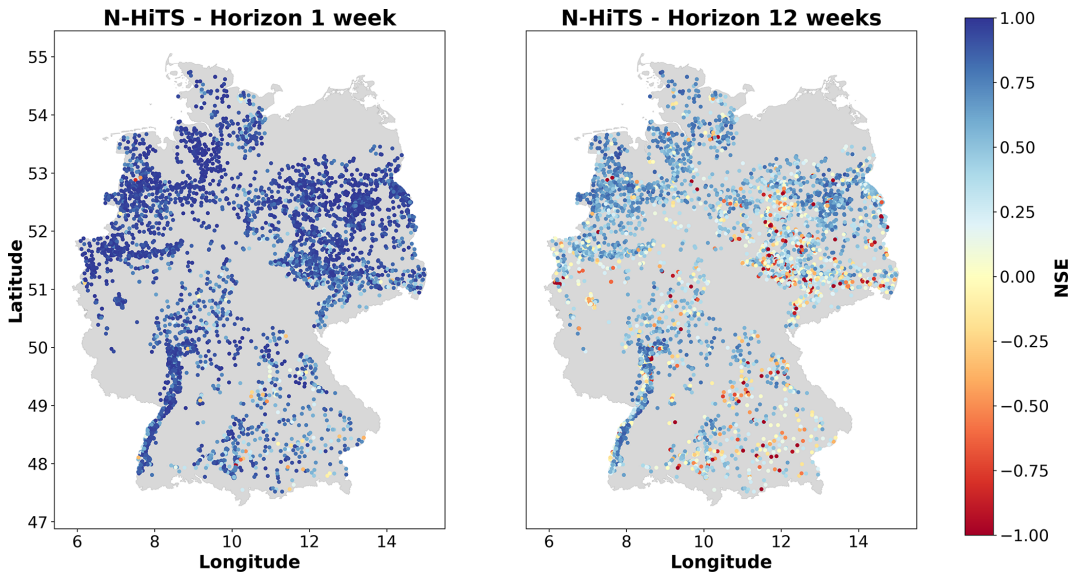

Figure 5. Performance of the N-HiTS model in terms of NSE across Germany for the 1-week prediction and 12-week prediction horizon. Results for the TFT model are shown in Fig. (B2).

the North and Central German Unconsolidated Rock District (northwestern Saxony), possibly lead to difficulties in predicting groundwater level fluctuations despite the high density of monitoring wells. For instance, Schroeter and Gläßer (2011) report a decline of the groundwater table of up to 70 m during the period of active mining in some parts of this area. Moreover, comparatively larger errors at certain locations and the erratic spatial pattern of model performance suggest that there are influencing factors not covered by the input features used; these could include anthropogenic influences such as water withdrawals. Filtering out monitoring wells with strong anthropogenic influences is not trivial for such a large study area, as most wells will exhibit some form of "unnatural" influence. Future studies should aim to identify time series features that indicate strong anthropogenic influences, which can then be used to filter or cluster for monitoring wells suitable for groundwater level prediction when information on water abstraction is not available.

4.2 Hydrogeological and environmental drivers of model performance

The correlation analysis of the NSE values with static features and time series features revealed weak or no correlations. The analysis of model performance across the different static feature categories indicated a few environmental conditions for which a better performance could be achieved but did not show pronounced differences.

The highest identified correlations suggest that the ML models' performance tended to improve when groundwater dynamics followed the expected seasonality, at monitoring wells with higher TWI values, at monitoring wells located at lower elevations, at monitoring wells that were surrounded by fewer or a less diverse set of landforms, and when groundwater hydrographs had a lower flashiness. The first four factors potentially characterise monitoring wells in lowlands, as TWI values are higher at lower elevations ($\rho = -0.43$), probably with a low depth to groundwater, enabling a distinct seasonality. These findings align with Gomez et al. (2024), who also found that model accuracy improved in areas with higher TWI and decreased in hilly regions. Examples of individual monitoring wells with particularly poor 12-week predictions where the expected seasonality was distorted (i.e. highest groundwater level in March, lowest in September) and with high flashiness are shown in Appendix B (Fig. B7).

For monitoring wells that deviate from the expected seasonality and exhibit lower-frequency variations, i.e. inert groundwater hydrographs, choosing a longer sequence length could improve predictions. However, for many of the monitoring wells there was a lack of sufficiently long time series to support longer sequence lengths and accordingly longer validation and test periods. Nevertheless, while we found that the ML models perform well for monitoring stations with an expected seasonal pattern, there are also instances where the models effectively capture unusual dynamics, such as hydrographs with a weak seasonality and extended phase of recession or a downward trend (Fig. B8).

The analysis of model performance across the categorical static features implied a better model performance in hydrogeologically relatively homogeneous areas (i.e. major hydrogeological districts) where many wells are located. Moreover, model performance tended to be better in porous aquifers and at sites with medium to moderate permeability coefficients. This finding aligns with theoretical expectations that in porous aquifers and at moderate permeabil-

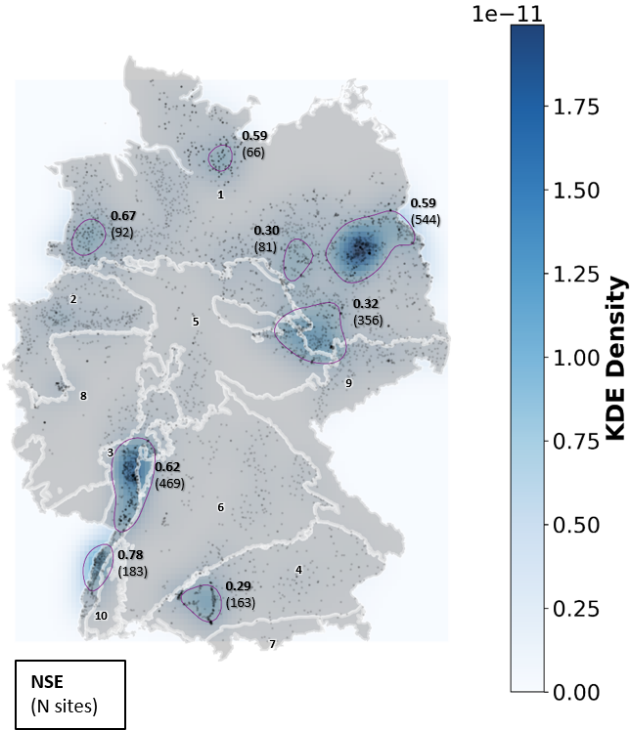

Figure 6. KDE of the density of monitoring wells. Areas with high data density are depicted in darker blue and encircled in purple. Median NSE values are reported in bold for each area with high data density. Dots represent the 5288 monitoring wells. The white lines and numbers from 1–10 represent the major hydrogeological districts. Legend (see Fig. B5): (1) North and Central German Unconsolidated Rock District, (2) Rhenish-Westphalian Lowland, (3) Upper Rhine Graben with Mainz Basin and North Hessian Tertiary, (4) Alpine Foreland, (5) Central German Fault-Block Land, (6) West and South German Scarplands and Fault-Block Land, (7) Alps, (8) West and Central German Basement, (9) Southeast German Basement, (10) Southwest German Basement.

ities groundwater flow is relatively slow and uniform and thus more predictable than in other hydrogeological systems such as karst and fractured aquifers, which exhibit highly anisotropic subsurface conditions with high flow velocities and heterogeneous groundwater dynamics (Bakalowicz, 2005; Hermans et al., 2023). Furthermore, this finding is also likely related to the high number of training examples for these conditions. Two additional environmental conditions with higher NSE values stood out: monitoring wells located in closed forests with deciduous broadleaf trees and monitoring wells located in peatlands. Wells in closed forests with deciduous broadleaf trees were primarily located in an area with high data density, the Upper Rhine Graben, suggesting that data availability rather than the presence of closed forests explains the good performance. In peatlands hydraulic conductivity is generally assumed to be low in the deeper peat layers (Kværner and Snilsberg, 2011). Additionally, high water tables may prevail because of limited lateral flow in areas with a low topographic gradient like northern Germany were most of the monitoring wells in peatlands are located (Oosterwoud et al., 2017). Thus, groundwater dynamics were more controlled and potentially mainly contingent on the climatic signal, which can explain the good predictive performance. In fact, the monitoring wells located in peatlands exhibited also relatively high values for the time series feature seasonal behaviour (median 0.59).

Overall, some hydrogeological and environmental conditions could be identified under which a good model performance could be achieved, in line with the understanding of the hydrogeological system, but the heterogeneous results also show the complexity of the data at hand.

4.3 The role of static features in global machine learning models for groundwater level prediction

Despite the relatively high feature importance scores for some static features, they contributed to improved predictive performance in only a small number of monitoring wells, suggesting limited utility for generalisation across data points with similar static feature levels. Previous studies demonstrated a stronger increase in predictive performance of global ML models when incorporating static features (Heudorfer et al., 2024; Li et al., 2022). However, these studies were conducted for a much smaller number of monitoring wells and the authors suggest that their models used the static features primarily as unique identifiers (Heudorfer et al., 2024; Li et al., 2022). One possible explanation for the addition of static features yielding only a small increase in the overall model performance is that the ML models used may have primarily learned from the dynamic input features. In contrast to the cited studies, the N-HiTS and TFT models were provided with the target variable (groundwater level) as an input feature, which was unexpectedly the most important input feature based on the feature importance of the TFT. It is important to note that by using a validation set and various techniques such as dropout and early stopping to avoid overfitting, the models were prevented from simply replicating historical groundwater levels. Moreover, in comparison with the model using solely dynamic input features, the model variants incorporating additional static features were partly "confused" by this information. While for few monitoring wells an improvement in model performance with the addition of static features could be achieved, e.g. for N-HiTS for the 12-week prediction 5.4 % of monitoring wells with an NSE of 0.5 saw an increase in model performance of 0.1 NSE or greater, for a small portion of monitoring wells the model performance declined. To be precise, 1.9 % of monitoring wells with an NSE of at least 0.5 exhibited a decrease in model performance of 0.1 NSE or greater. Possibly, the wells exhibiting a decline in performance had groundwater dynamics that differed strongly from other monitoring wells with similar static feature values.

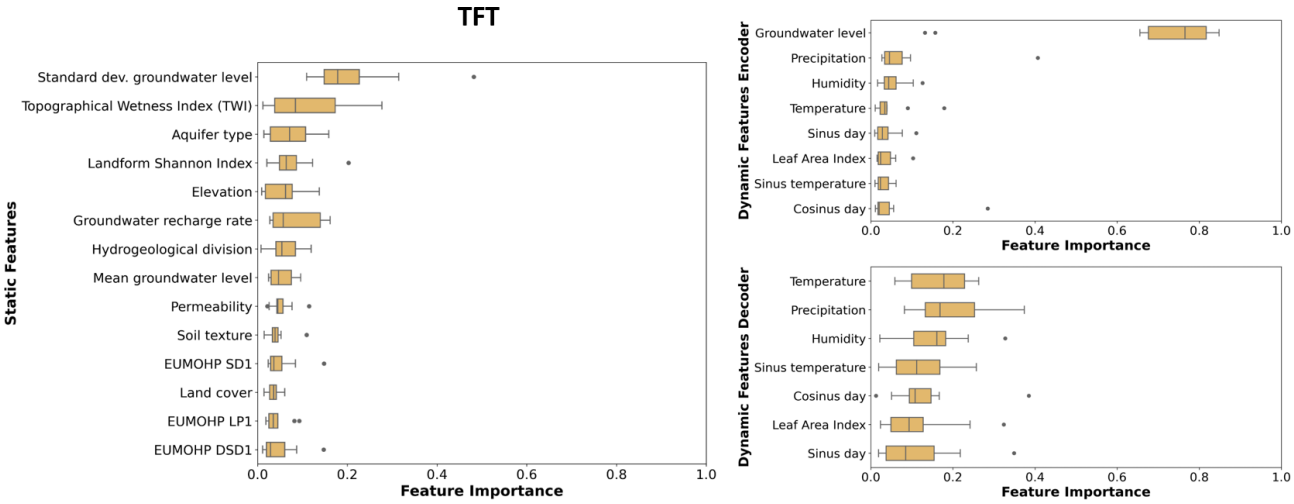

Figure 7. Variable importance of the TFT model trained with all static and dynamic input features. Each box plot shows 10 importance values based on the different weight of initialisations. Variables are ranked based on their median importance. Results for the other TFT model variants are shown in Appendix B (Fig. B9).

Further reasons why the models may have been unable to extract meaningful relationships between the groundwater level and the static features may be due to their inherent uncertainties, which make it challenging for the models to use them to generalise to wells with similar static feature levels. These uncertainties stem from (1) the fact that the best spatial aggregation that accurately reflects the static input features and their relation to the groundwater level depends on the specific feature and monitoring location and may not be localised to the pixel level (1 km buffer) at the surface and (2) the regionalised nature of these features, wherein values are derived by extrapolating from discrete points into broader spatial contexts. In order to improve future groundwater level predictions, it could be beneficial to utilise higher-resolution maps or combine two-dimensional maps with time series data (e.g. Addimando et al., 2022), as well as to test different spatial aggregations of static features. The features employed in this study predominantly describe surface or near-surface conditions (e.g. soil texture), since they are easily accessible and available for a large area. Groundwater flow dynamics are strongly affected by subsurface factors and the underlying complexity of the hydrogeology. Thus, the incorporation of features that describe subsurface properties, such as the depth and thickness of the aquifer in the vicinity of the monitoring wells, could enhance the predictive performance. However, at present, information on such features is often not extensively available.

Lastly, the marginal improvement in model performance for the N-HiTS model with static features might be attributed to differences in the data fusion of static and dynamic input features compared to the TFT architecture. N-HiTS uses an additional stack composed of fully connected layers for exogenous variables (FCNN), while the TFT uses GRN encoders that provide context for the temporal processing.

4.4 Feature importance and important past time steps

The TFT offers an intrinsic feature importance method through the VSNs, providing insights into which features are relevant for prediction. This makes this architecture attractive, as no post-analysis of feature importance was required, such as Shapley values or permutation importance. Nevertheless, this study demonstrated that the results should be interpreted with caution, as some static features exhibited relatively high importance values (standard deviation of groundwater level, TWI) yet did not enhance the overall predictive performance of the TFT. Moreover, the feature importance values of the TFT had a high dependency on the model weight initialisation, which underpins the necessity of running several initialisations when using this approach. The observed variability in feature importance for the features day of the year, temperature, and sinus of the temperature can be attributed to the fact that these features provide the model with information on the yearly seasonality. If one of them becomes important, the importance of the other features tends to diminish.

The TFT architecture also offers insights into the most important past time steps through the attention scores. The attention scores reflected the seasonality in the data, which is also represented in the autocorrelation function (Fig. B12), and highlighted its importance as well as the importance of the immediate past for groundwater level prediction. That attention scores from a TFT model mainly reflect seasonality in hydrological or hydrogeological data was also observed in a study on streamflow prediction (Koya and Roy, 2023). Future studies may explore using longer sequence lengths to

capture longer range dependencies, wherein attention layers can reveal the relevance of these time points, though such experiments likely come with higher computational costs.

5 Conclusion

In this study, two recently developed global ML architectures, TFT and N-HiTS, were applied to predict groundwater levels for up to 12 weeks for 5288 monitoring wells distributed across Germany. The results demonstrated that these architectures are suitable for seasonal groundwater level prediction for thousands of monitoring wells, enabling good predictions across a large area with only one model, whereby the N-HiTS model showed better predictive performance for our dataset. While overall predictive performance was good the models failed to accurately predict groundwater levels for a considerable number of wells (about 15 % with an NSE below 0 with the best-performing model), which highlights the hydrogeological complexity of the study area and the possible influence of anthropogenic factors such as water withdrawals that could not be captured by the input features. To address this issue, future research should aim at filtering out monitoring wells with strong anthropogenic influences, for example, based on suitable time series features or, if possible, include features that describe anthropogenic factors such as water abstractions.

In contrast to our expectations, the addition of static features in the models resulted in improved performance only for a small number of monitoring wells. Regarding the overall performance across all 5288 monitoring wells, the static features marginally improved the overall performance of the N-HiTS model, but not for the TFT model. Future studies could examine the potential benefits of alternative representations of static features, such as two-dimensional maps or more detailed information on subsurface processes, in order to more accurately reflect the true conditions at and around a monitoring well. Additionally, testing and using different buffer sizes, besides the 1 km radius used in our study, for data extraction around the monitoring wells could lead to further improvements in the use of static features. Furthermore, the impact of static features on the model performance may be more pronounced in exogenous-only models that do not use the groundwater level as an input feature, arguably the most important predictive feature.

The correlations between model performance and different static features, as well as time series features, were found to be low. This may again be attributed to the complex hydrogeological conditions across Germany, as well as potential anthropogenic influences. Nevertheless, the results of the correlation analysis implied that more accurate predictions can be made in lowland areas with less complex surrounding landscapes and groundwater hydrographs exhibiting a seasonal pattern. Moreover, comparisons of model performance across the categorical static features pointed out that a slightly better performance can be obtained in porous aquifers, areas with medium to moderate permeability coefficients, and in peatlands. Areas with a high data density appear to facilitate more accurate predictions, which aligns with theoretical expectations that high-capacity ML models benefit from a large number of training examples with a high data diversity (Nearing et al., 2021; Kratzert et al., 2024).

The feature importance obtained from the TFT showed a relatively high variability across the 10 initialisations. This should be taken into account when utilising the model interpretability insights provided by the TFT, or alternatively, model-agnostic techniques should be employed. Nevertheless, the results indicated that the TFT mainly learned from the provided groundwater levels and the standard deviation of the groundwater levels. Other important dynamic features were precipitation, humidity, and temperature, while a notable important static feature was the TWI.

In conclusion, the state-of-the-art global ML architectures TFT and N-HiTS represent valuable additions to the thus far established LSTM architecture for groundwater level prediction. The N-HiTS architecture demonstrated superior predictive capabilities compared to the TFT architecture based on the predictive performance for the compiled dataset. This makes it a powerful ML architecture for groundwater level predictions in large, hydrogeologically complex areas.

Appendix A: Additional information on training data and model evaluation

A1 Length of training period

Table A1. Descriptive statistics on the length of the training period.

	Length training period (y)
No. monitoring wells	5288
Mean	16.42
SD	4.68
Min	6
25 %	13
50 %	20
75 %	20
Max	20

A2 Time series characteristics of the groundwater level time series in training, validation, and test period

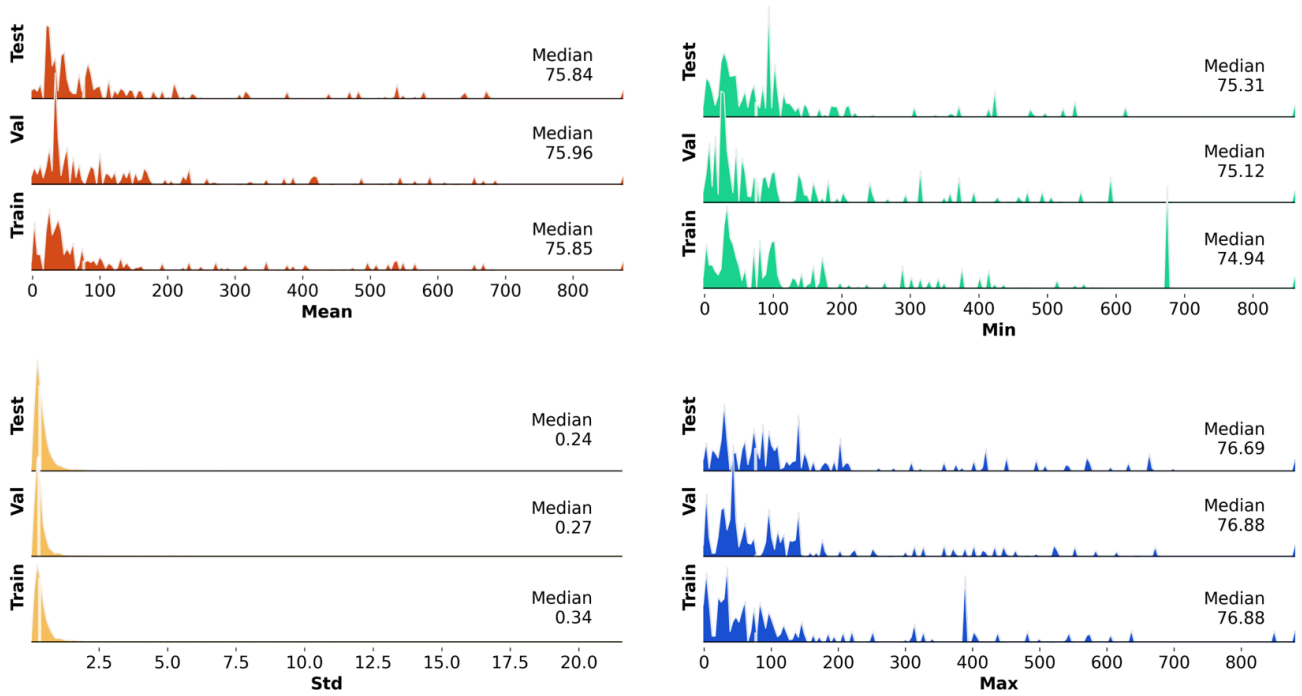

Figure A1. Distribution of the mean, standard deviation (SD), minimum (min), and maximum (max) of the 5288 groundwater level time series for the training, validation, and test period.

A3 Description of relative mean bias error and interval score

As additional error metrics the relative mean bias error (rMBE) and the interval score for evaluating probabilistic forecasts are reported. The rMBE is defined as the mean deviance divided by a constant, for which the standard deviation of the groundwater levels is chosen. Negative rMBE values indicate underestimation, positive values indicate overestimation, and a value close to zero can indicate a good estimate. Nevertheless, a value of zero does not necessarily indicate a perfect estimate, since under- and overestimation can cancel each other out. Hence, the rMBE is only sensitive to biases having a dominant direction. It is defined as follows:

$$\mathrm{rMBE} = \frac{1}{\mathrm{SD}(\boldsymbol{y})} \frac{\sum_{t=1}^{T}(\hat{y}_t - y_t)}{T}. \tag{A1}$$

The interval score (Eq. A2) combines two properties of a prediction interval: (1) the sharpness, which refers to the concentration of the predictive distribution and is a property of the prediction only, and (2) the calibration, which denotes how much the true value is outside of the predicted interval. In the equation t denotes a time step in the time series, T denotes the total number of time steps, y_t denotes the observed value of the target variable at time step t, $\hat{y}_t$ denotes the estimated value at time step t, and $\hat{y}_t^{\top}$ and $\hat{y}_t^{\perp}$ denote the upper and lower bound of the predicted distribution at time step t. Smaller values of the interval score indicate a better estimate and a narrower prediction interval. A normalised version of the interval score was used, obtained by dividing the interval score by the standard deviation of the true values.

$$\mathrm{rIS} = \frac{1}{\mathrm{SD}(\boldsymbol{y})} \times \frac{\sum_{t=1}^{T}\left(\overbrace{(\hat{y}_t^{\top} - \hat{y}_t^{\perp})}^{\text{sharpness}} + \overbrace{\frac{2}{\alpha}\left(\max(\hat{y}_t^{\perp} - y_t, 0) + \max(y_t - \hat{y}_t^{\top}, 0)\right)}^{\text{calibration}}\right)}{T} \tag{A2}$$

Appendix B: Model performance and model interpretability

B1 Model performance comparison across prediction horizons

Table B1. Median NSE values for the prediction of groundwater levels of the 5288 wells for each horizon. Reported values are based on the median NSE of the 10 initialisations for each well. P values refer to the difference between the model variants provided with all input features compared to the purely dynamic model (one-sided Mann–Whitney U test).

Horizon	Median NSE TFT	Median NSE TFT (purely dynamic)	P value TFT	Median NSE N-HiTS	Median NSE N-HiTS (purely dynamic)	P value N-HiTS
1	0.91	0.91	3.22×10^{-3}	0.92	0.92	8.89×10^{-2}
2	0.85	0.85	5.71×10^{-2}	0.87	0.86	3.91×10^{-2}
3	0.80	0.80	1.29×10^{-2}	0.82	0.82	1.83×10^{-2}
4	0.75	0.74	5.76×10^{-3}	0.79	0.78	4.84×10^{-3}
5	0.69	0.69	2.56×10^{-3}	0.74	0.73	3.01×10^{-3}
6	0.64	0.63	2.53×10^{-3}	0.70	0.68	2.30×10^{-3}
7	0.59	0.57	3.69×10^{-3}	0.66	0.64	6.66×10^{-4}
8	0.54	0.53	2.00×10^{-2}	0.62	0.61	5.20×10^{-4}
9	0.48	0.48	2.34×10^{-2}	0.59	0.57	2.05×10^{-5}
10	0.44	0.43	3.51×10^{-2}	0.55	0.53	5.84×10^{-5}
11	0.39	0.39	2.59×10^{-1}	0.53	0.51	3.18×10^{-5}
12	0.34	0.35	4.21×10^{-1}	0.50	0.47	7.42×10^{-6}

Table B2. Median interval scores for TFT and N-HiTS models for each horizon. Reported values are based on the median interval score of the models provided with all input features compared to the purely dynamic model. P values refer to the difference between the model variants and the purely dynamic model (two-sided Mann–Whitney U test).

Horizon	Median interval score TFT	Median interval score TFT (purely dynamic)	P value TFT	Median interval score N-HiTS	Median interval score N-HiTS (purely dynamic)	P value N-HiTS
1	1.54	1.62	8.634×10^{-24}	1.19	1.26	1.659×10^{-7}
2	1.98	2.14	4.029×10^{-43}	1.86	1.70	1.698×10^{-25}
3	2.51	3.15	1.192×10^{-286}	2.25	2.19	7.403×10^{-3}
4	2.86	3.01	7.466×10^{-28}	2.65	2.44	1.694×10^{-32}
5	3.29	3.44	2.533×10^{-15}	3.01	2.88	9.202×10^{-15}
6	3.65	4.08	3.570×10^{-111}	3.16	3.16	1.982×10^{-1}
7	3.72	4.07	4.571×10^{-63}	3.62	3.41	9.869×10^{-24}
8	4.13	4.56	2.129×10^{-109}	3.65	3.69	6.766×10^{-1}
9	4.81	4.53	2.223×10^{-59}	3.82	3.96	3.939×10^{-9}
10	4.44	4.82	2.233×10^{-68}	3.84	3.90	1.114×10^{-1}
11	4.53	4.61	3.054×10^{-3}	3.99	3.80	1.023×10^{-17}
12	4.45	4.41	4.126×10^{-8}	3.89	3.68	2.131×10^{-17}

B2 Performance comparison with the naïve forecast

Table B3. Median NSE values for the prediction of groundwater levels of the 5288 wells with the naïve forecast.

Horizon	Median NSE naïve forecast
1	0.91
2	0.82
3	0.71
4	0.60
5	0.47
6	0.34
7	0.20
8	0.06
9	−0.10
10	−0.25
11	−0.39
12	−0.52

Table B4. Comparison of the number of monitoring wells where ML architectures or naïve forecast perform better depending on the model type for the 1-week prediction horizon.

Model type	*N* wells better with ML	*N* wells better with naïve forecast
N-HiTS (purely dynamic)	449	7
N-HiTS	515	4
TFT (purely dynamic)	450	42
TFT	426	39

Table B5. Comparison of flashiness values for groundwater level time series where ML architectures improved predictions by at least 0.1 NSE over the naïve forecast (called subgroup in the table; see also Table B4). *P* values were obtained with a one-sided Mann–Whitney *U* test by comparison against flashiness values of all groundwater level time series.

Model type	Subgroup mean flash.	Subgroup median flash.	Overall mean flash.	Overall median	*P* value
N-HiTS (purely dynamic)	0.22	0.15	0.12	0.08	1.94×10^{-82}
N-HiTS	0.23	0.16	0.12	0.08	8.15×10^{-96}
TFT (purely dynamic)	0.22	0.15	0.12	0.08	1.01×10^{-65}
TFT	0.22	0.15	0.12	0.08	4.69×10^{-67}

B3 Prediction intervals

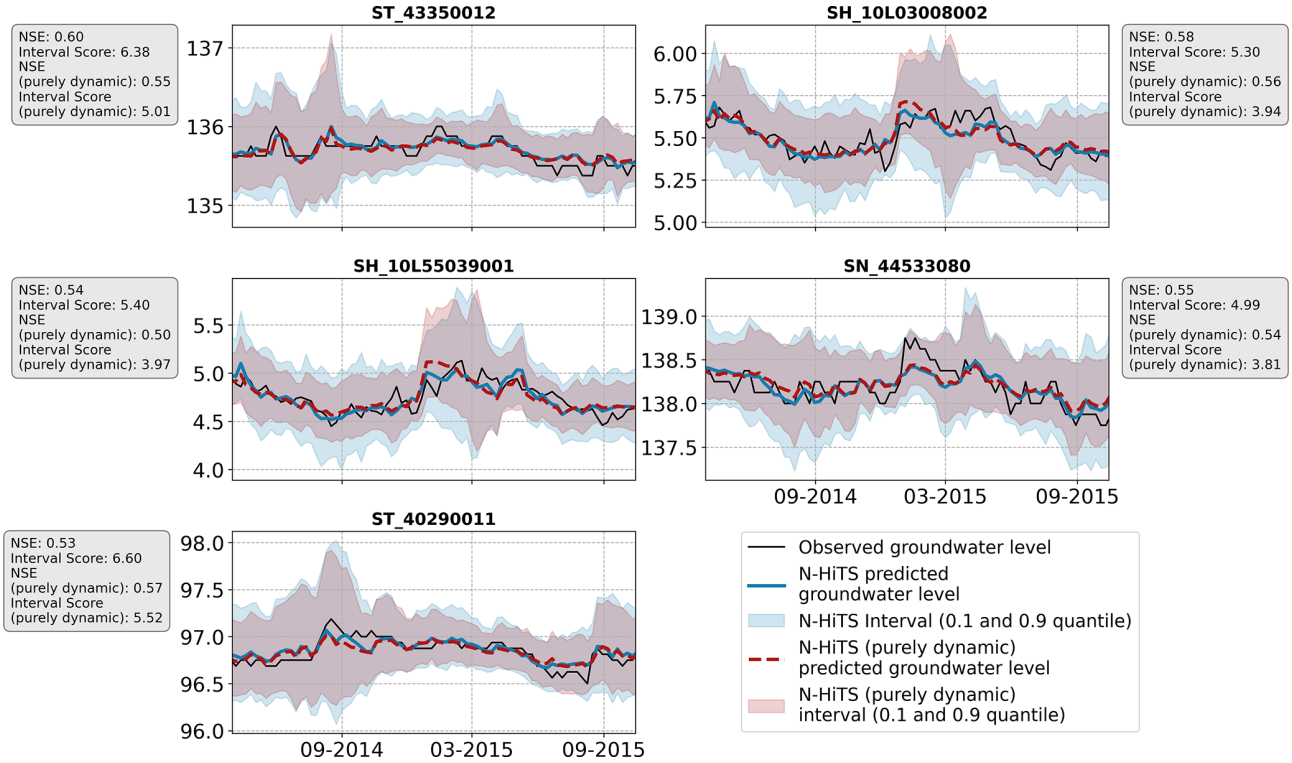

Figure B1. Five example hydrographs with narrower prediction intervals obtained with the N-HiTS (purely dynamic) model than the N-HiTS model provided with static and dynamic features.

B4 Model performance of TFT model trained without historic groundwater level observations

Table B6. Performance summary of the TFT model variant trained without historic groundwater level data as input, i.e. a model only trained on exogenous features. Two variants have been trained: with exogenous static and dynamic features and an ablated version with exogenous dynamic features.

Model type	Horizon	NSE	RMSE	rMBE	Interval score
TFT	1	0.37	0.16	0.22	3.72
TFT	12	0.08	0.18	0.26	4.49
TFT (purely dynamic)	1	0.22	0.18	0.23	3.95
TFT (purely dynamic)	12	−0.04	0.20	0.28	4.57

B5 Spatial patterns of model performance

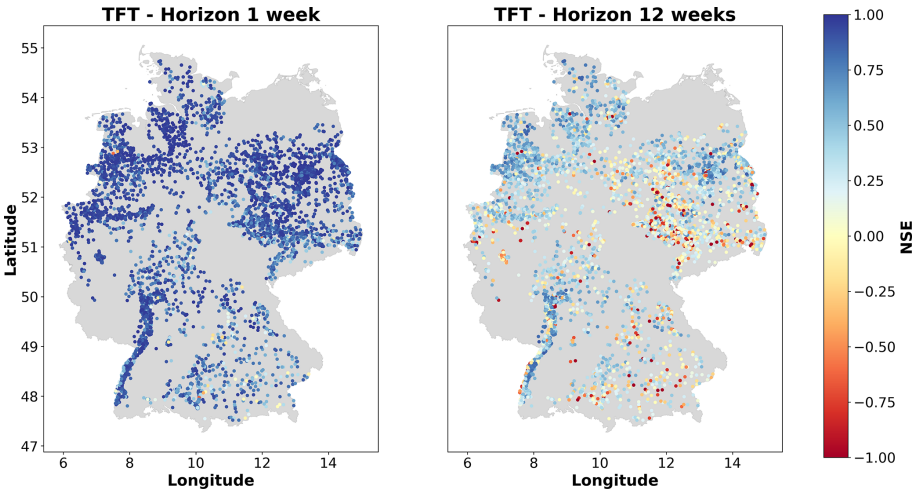

Figure B2. Spatial distribution of the NSE values for the TFT model provided with static and dynamic features across Germany for the 1-week prediction and 12-week prediction horizon.

B6 Hydrogeological and environmental drivers of model performance

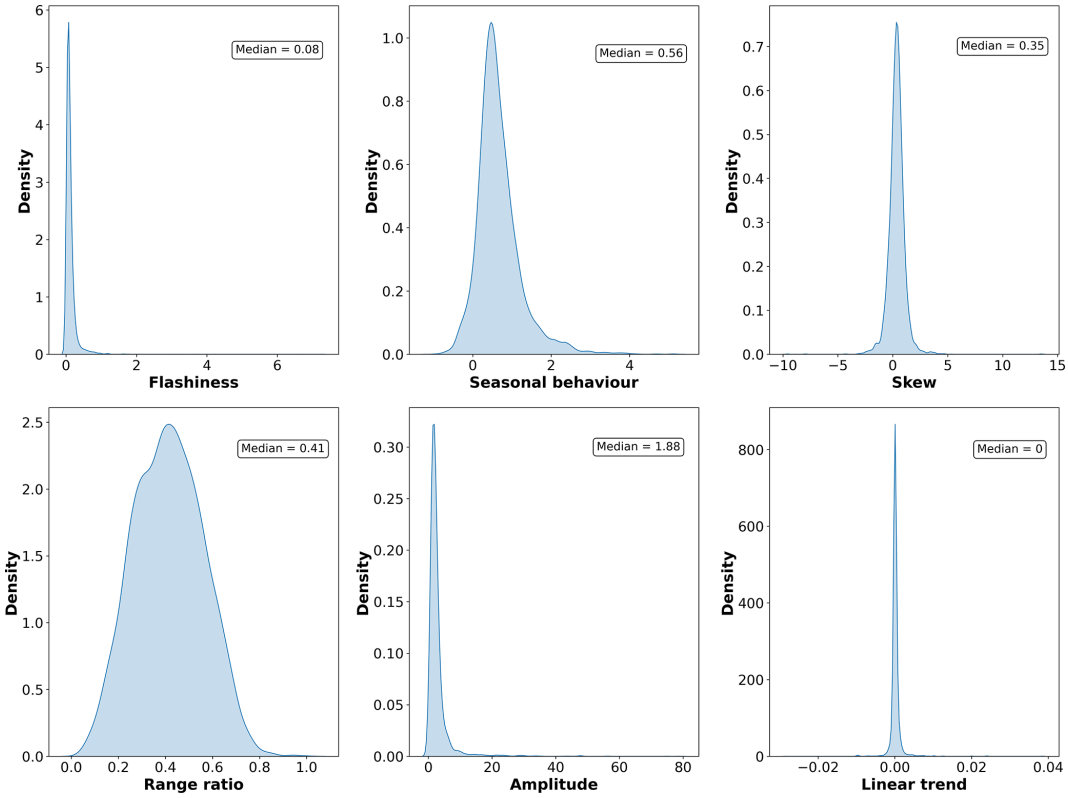

Figure B3. Distributions of the calculated time series feature values for the 5288 monitoring wells.

Table B7. Median NSE values for different static features for all model variants and a prediction horizon of 12 weeks. n is the number of wells in each major hydrogeological district. The legend for the major hydrogeological districts is shown in Fig. B5.

Static feature	n	Median NSE N-HiTS	Median NSE N-HiTS (purely dynamic)	Median NSE TFT	Median NSE TFT (purely dynamic)
Aquifer type					
Fractured and karstified aquifers	185	0.45	0.44	0.33	0.33
Fractured and porous aquifers	95	0.37	0.36	0.27	0.26
Fractured aquifers	612	0.39	0.36	0.27	0.27
Porous aquifers	4395	0.52	0.49	0.36	0.36
Land cover					
Closed forest, deciduous broadleaf	187	0.6	0.56	0.39	0.43
Closed forest, evergreen needleleaf	463	0.55	0.48	0.35	0.36
Closed forest, mixed	129	0.55	0.48	0.37	0.42
Cultivated and managed vegetation/agriculture	3611	0.51	0.49	0.36	0.35
Herbaceous vegetation	65	0.34	0.32	0.23	0.2
Open forest, mixed	3	0.42	0.57	0.42	0.52
Open forest, unknown	15	0.06	−0.22	−0.29	−0.17
Urban/built-up	792	0.4	0.37	0.29	0.31
Major hydrogeological district					
1	2564	0.56	0.54	0.4	0.4
2	387	0.47	0.46	0.36	0.35
3	807	0.64	0.62	0.54	0.52
4	422	0.3	0.24	0.17	0.14
5	441	0.32	0.31	0.17	0.19
6	265	0.49	0.44	0.35	0.36
7	12	0.26	0.26	0.14	0.14
8	164	0.35	0.34	0.22	0.29
9	211	0.22	0.18	0.08	0.11
10	15	0.7	0.68	0.58	0.56
Permeability coefficient					
High ($> 1 \times 10^{-3}$–1×10^{-2})	788	0.37	0.33	0.22	0.21
Highly variable	110	0.29	0.26	0.24	0.21
Low to extremely low ($< 1 \times 10^{-5}$)	520	0.38	0.36	0.26	0.25
Medium to moderate ($> 1 \times 10^{-5}$–1×10^{-3})	3600	0.56	0.53	0.4	0.4
Moderate to low ($> 1 \times 10^{-6}$–1×10^{-4})	265	0.33	0.29	0.24	0.23
Soil texture					
Clay loams	592	0.47	0.48	0.38	0.39
Clay silts	876	0.38	0.36	0.25	0.28
Loams	127	0.32	0.3	0.17	0.08
Loamy sands	1376	0.57	0.54	0.42	0.41
Loamy silts	287	0.38	0.34	0.2	0.24
Open cast mining sites	7	0.19	0.27	0.14	0.19
Peatlands	286	0.66	0.63	0.46	0.44
Pure sands	807	0.6	0.58	0.43	0.43
Sandy loams	558	0.44	0.38	0.29	0.27
Silty clays	54	0.5	0.46	0.36	0.29
Silty sands	19	0.29	0.38	0.24	0.18
Urban settlements	287	0.35	0.33	0.26	0.32

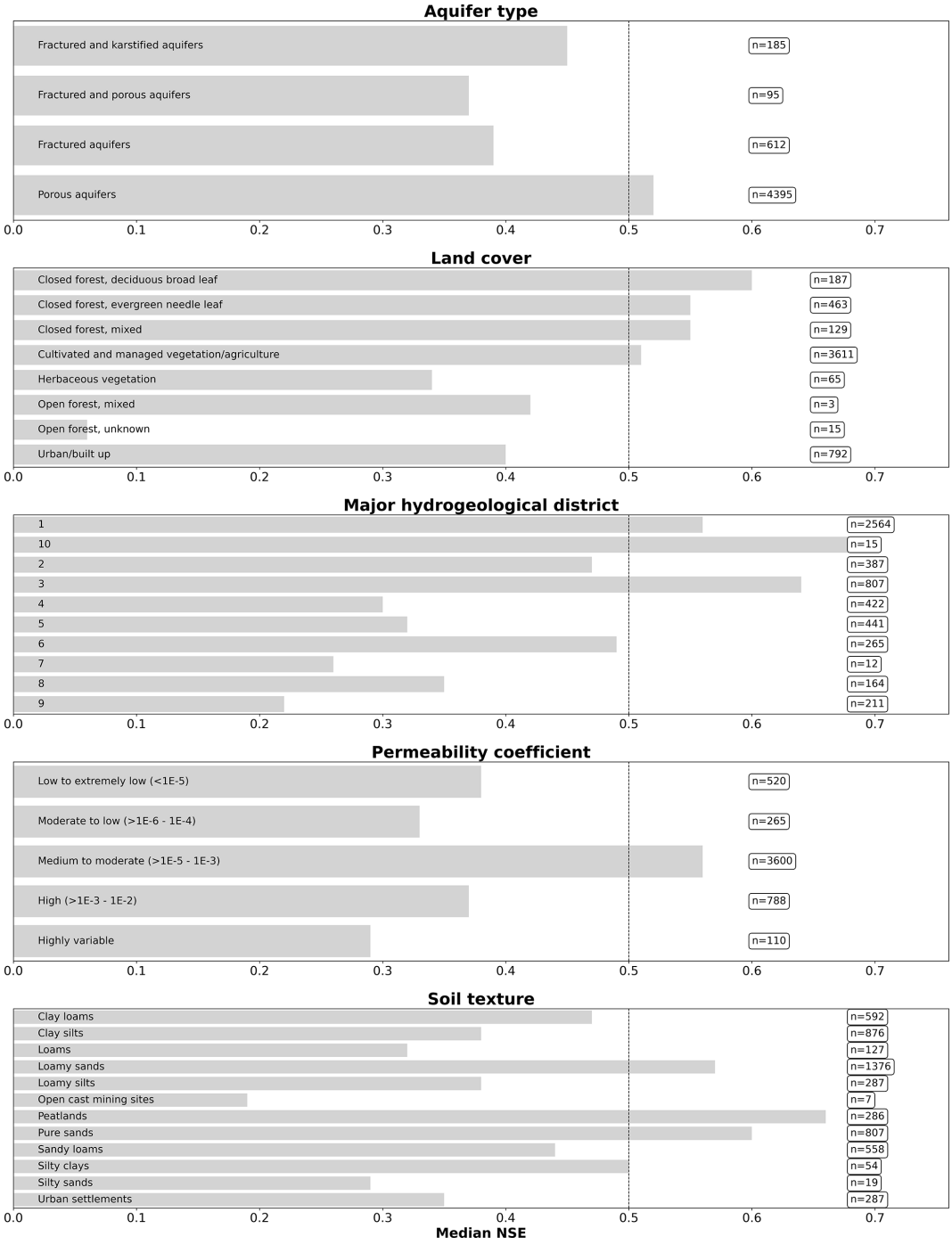

Figure B4. Bar plots showing the median NSE per static feature category for the N-HiTS model provided with static and dynamic features for the 12-week prediction horizon. The categories with the highest values are highlighted. The sample size is given behind each bar. The dotted vertical line indicates the overall NSE for this model variant (0.5). The legend for the major hydrogeological districts is in Fig. B5.

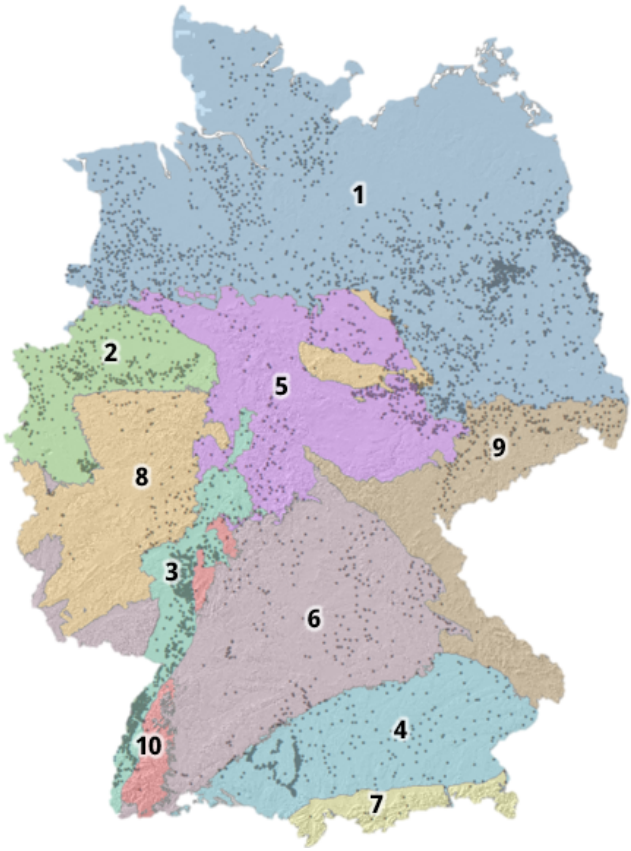

Figure B5. The major hydrogeological districts for Germany. Dots represent the 5288 monitoring wells. Legend: (1) North and Central German Unconsolidated Rock District, (2) Rhenish-Westphalian Lowland, (3) Upper Rhine Graben with Mainz Basin and North Hessian Tertiary, (4) Alpine Foreland, (5) Central German Fault-Block Land, (6) West and South German Scarplands and Fault-Block Land, (7) Alps, (8) West and Central German Basement, (9) Southeast German Basement, (10) Southwest German Basement.

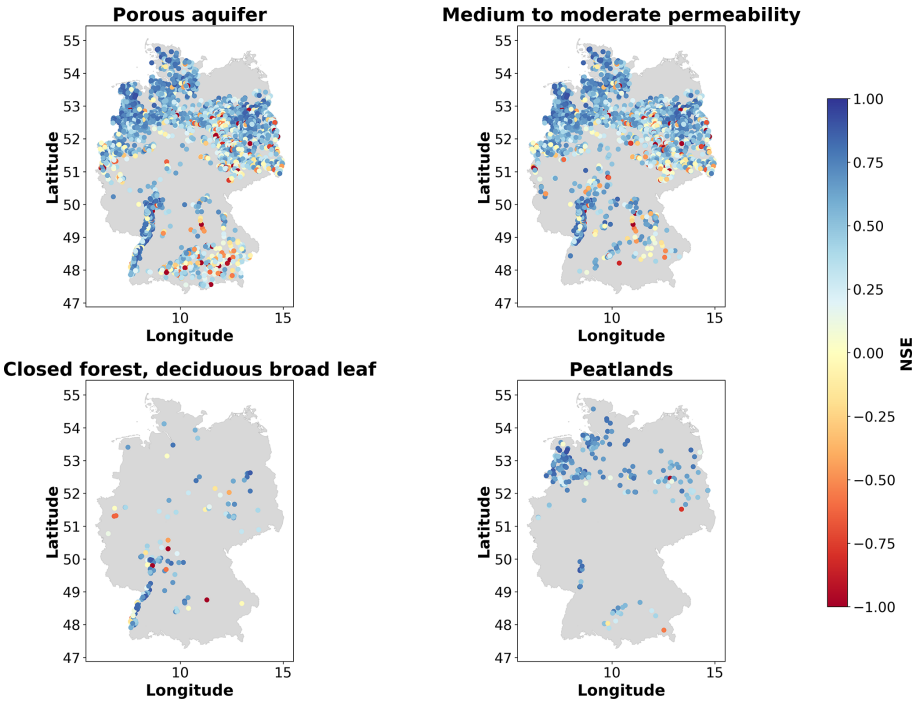

Figure B6. Locations of the monitoring wells with the highest median NSE values for the N-HiTS model provided with static and dynamic features for the 12-week prediction horizon for monitoring wells located in porous aquifers, at sites with medium permeabilities, in closed forest with deciduous broadleaf, and in peatlands.

B7 Predictions for individual wells

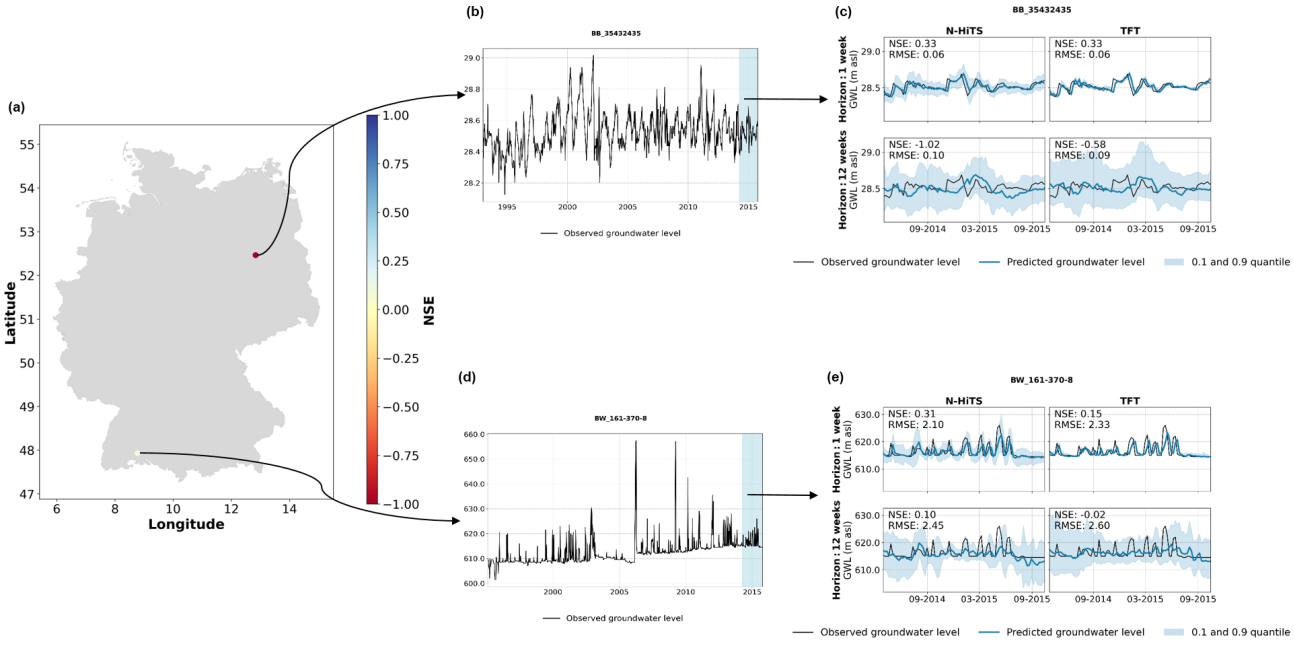

Figure B7. Examples of two monitoring wells with groundwater hydrographs with a distorted seasonality (BB_35432435) and a high flashiness (BW_161-370-8). Panels **(b)** and **(d)** show the hydrographs for the entire period of training, validation, and test (the blue-shaded area marks the test period). Panels **(c)** and **(e)** show the predictions for the N-HiTS and TFT model provided with static and dynamic input features for the test period.

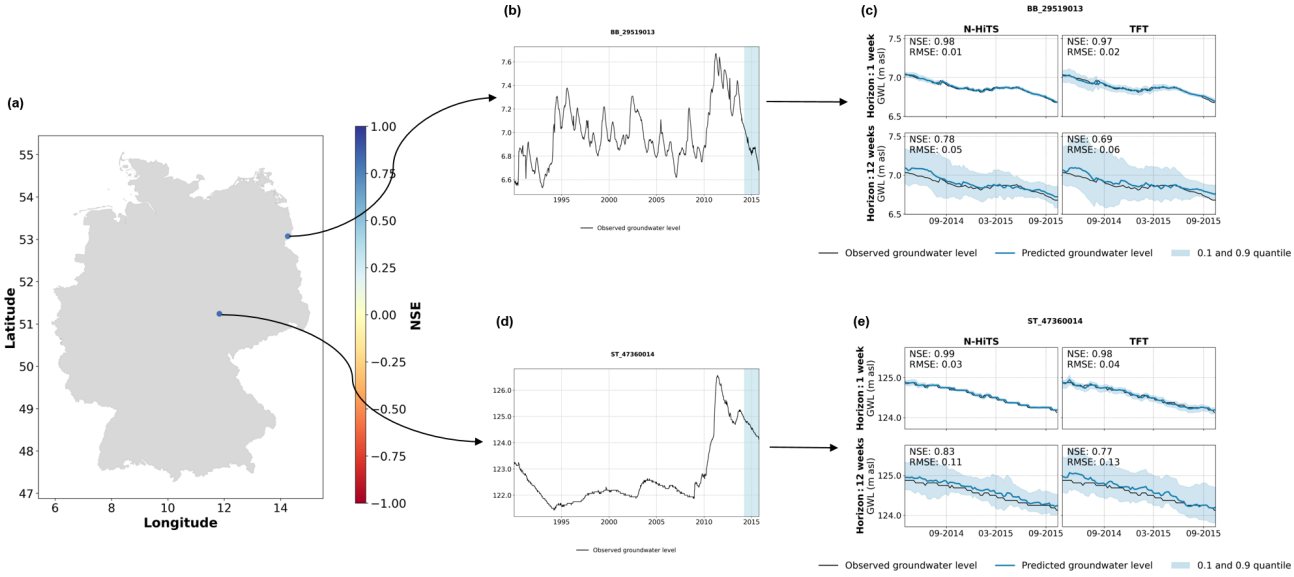

Figure B8. Examples of two monitoring wells with groundwater hydrographs with inert groundwater dynamics where the models achieved good predictive performance for the 12-week prediction horizon. Panels **(b)** and **(d)** show the hydrographs for the entire period of training, validation, and test (the blue-shaded area marks the test period). Panels **(c)** and **(e)** show the predictions for the N-HiTS and TFT model provided with static and dynamic input features for the test period.

B8 Feature importance of the TFT models

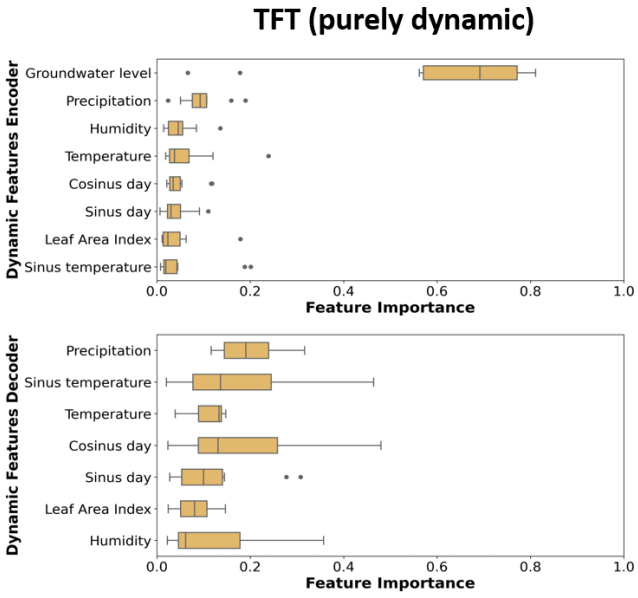

Figure B9. Variable importance values of the TFT (purely dynamic) model. Each box plot shows 10 importance values based on the different weight initialisations. Variables are ranked based on their median importance.

B9 Attention curves

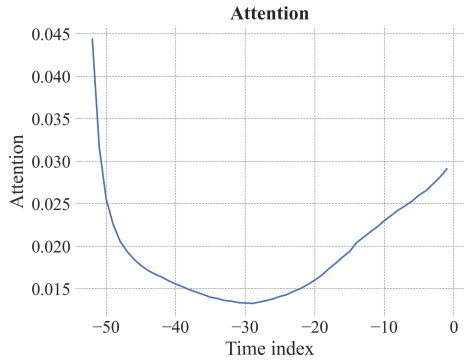

Figure B10. Attention curve for the 1-week prediction of the TFT provided with static and dynamic input variables highlighting important past time steps. Time index is in weeks. The autocorrelation function for 52 weeks is shown in Fig. B12.

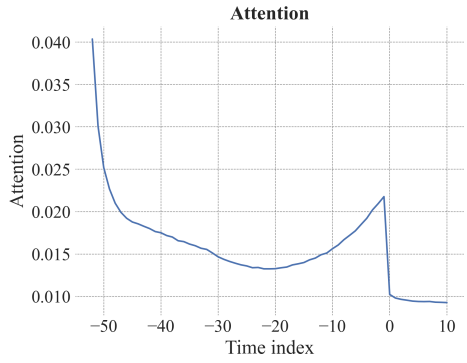

Figure B11. Attention curve for the 12-week prediction of the TFT provided with static and dynamic input variables highlighting important past time steps. Time index is in weeks. The autocorrelation function for 52 weeks is shown in Fig. B12.

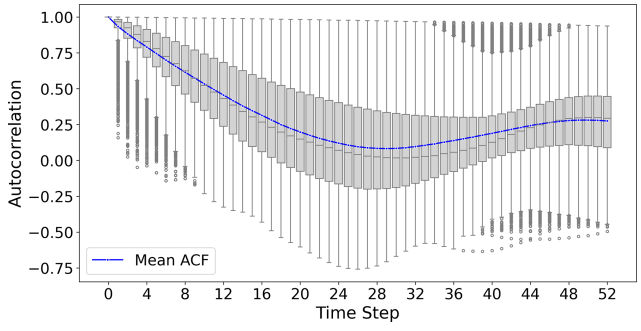

Figure B12. Box plots show the distribution of the autocorrelation function for 52 weeks across all monitoring wells. The blue line shows the mean autocorrelation function.

Code and data availability. The code necessary to reproduce our results is available on GitHub: https://github.com/KunzstBGR/global-groundwater-models-main (Kunz, 2025). The TFT and N-HiTS architectures were implemented and trained with Pytorch forecasting (Version 1.0.0) and Pytorch lightning (Version 2.0.4), which are open-source deep learning modules based on the Pytorch framework (Paszke et al., 2019). A complete list of all the Python modules used is available at the GitHub repository (requirements_GGWM.txt file). All groundwater level data are available free of charge from the respective local authorities upon request.

Author contributions. SK: conceptualisation, data curation, investigation, formal analysis, methodology, software, writing. AS: conceptualisation, data curation, software, methodology, writing. MW: conceptualisation, data curation, visualisation, methodology, writing (review and editing). MN: data curation, writing (review and editing). TC: software, methodology, writing (review and editing). FB: conceptualisation, resources, methodology, writing (review and editing). SB: conceptualisation, funding acquisition, resources, methodology, writing (review and editing).

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

Financial support. This research has been supported by the Bundesministerium für Bildung und Forschung (grant no. KIMoDIs FKZ 02WGW1662A).

Review statement. This paper was edited by Alberto Guadagnini and reviewed by Leonardo Sandoval, Abel Henriot, and one anonymous referee.

References

Addimando, N., Engel, M., Schwarz, F., and Batič, M.: A DEEP LEARNING APPROACH FOR CROP TYPE MAPPING BASED ON COMBINED TIME SERIES OF SATELLITE AND WEATHER DATA, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLIII-B3-2022, 1301–1308, https://doi.org/10.5194/isprs-archives-XLIII-B3-2022-1301-2022, 2022.

Althoff, D., Rodrigues, L. N., and Bazame, H. C.: Uncertainty quantification for hydrological models based on neural networks: the dropout ensemble, Stoch. Env. Res. Risk A., 35, 1051–1067, https://doi.org/10.1007/s00477-021-01980-8, 2021.

Amatulli, G., Domisch, S., Tuanmu, M.-N., Parmentier, B., Ranipeta, A., Malczyk, J., and Jetz, W.: A suite of global, cross-scale topographic variables for environmental and biodiversity modeling, Sci. Data, 5, 180040, https://doi.org/10.1038/sdata.2018.40, 2018.

Bakalowicz, M.: Karst groundwater: a challenge for new resources, Hydrogeol. J., 13, 148–160, https://doi.org/10.1007/s10040-004-0402-9, 2005.

Beven, K. J. and Kirkby, M. J.: A physically based, variable contributing area model of basin hydrology / Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant, Hydrol. Sci. B., 24, 43–69, https://doi.org/10.1080/02626667909491834, 1979.

BGR: Bodenarten in Oberböden Deutschlands 1 : 1.000.000 (BOART1000), Digitaler Datenbestand, version 2.0, UUID: 410B839B-DCF2-45BE-8788-3C65BBEC2D59, https://geoportal.bgr.de/mapapps/resources/apps/geoportal/index.html?lang=de#/datasets/portal/DADB8BB6-4A7A-4CB7-908D-EA0767B068D7 (last access: 14 October 2024), 2007.

BGR: Mittlere jährliche Grundwasserneubildung von Deutschland 1 : 1.000.000 (GWN1000), Digital map data, version 1, https://geoportal.bgr.de/mapapps/resources/apps/geoportal/index.html?lang=de#/datasets/portal/40E14FF1-99D4-43DA-AF7B-C039F0463BF8 (last access: 14 October 2024), 2019.

BGR and SGD: Hydrogeologische Übersichtskarte von Deutschland 1 : 250.000 (HÃœK250), Digitaler Datenbestand, version 1.0.3, UUID: afbc7573-d18d-453e-bff3-6477af048770, https://geoportal.bgr.de/mapapps/resources/apps/geoportal/index.html?lang=de#/datasets/portal/61ac4628-6b62-48c6-89b8-46270819f0d6 (last access: 14 October 2024), 2019.

Buchhorn, M., Smets, B., Bertels, L., De Roo, B., Lesiv, M., Tsendbazar, N. E., Linlin, L., and Tarko, A.: opernicus Global Land Service: Land Cover 100m: Version 3 Globe 2015–2019: Product User Manual, Zenodo, https://doi.org/10.5281/zenodo.3938963, 2020.

Cao, D., Wang, Y., Duan, J., Zhang, C., Zhu, X., Huang, C., Tong, Y., Xu, B., Bai, J., Tong, J., and Zhang, Q.: Spectral Temporal Graph Neural Network for Multivariate Time-series Forecasting, arXiv [preprint], https://doi.org/10.48550/arXiv.2103.07719, 2021.

Challu, C., Olivares, K. G., Oreshkin, B. N., Garza, F., Mergenthaler-Canseco, M., and Dubrawski, A.: N-HiTS: Neural Hierarchical Interpolation for Time Series Forecasting, arXiv [preprint], https://doi.org/10.48550/arXiv.2201.12886, 2022.

Chiaburu, T. and Bießmann, F.: Interpretable Time Series Models for Wastewater Modeling in Combined Sewer Overflows, Proced. Comput. Sci., 237, 155–162, https://doi.org/10.1016/j.procs.2024.05.091, 2024.

Chidepudi, S. K. R., Massei, N., Jardani, A., and Henriot, A.: Groundwater level reconstruction using long-term climate reanalysis data and deep neural networks, J. Hydrol.-Regional Studies, 51, 101632, https://doi.org/10.1016/j.ejrh.2023.101632, 2024.

Chidepudi, S. K. R., Massei, N., Jardani, A., Dieppois, B., Henriot, A., and Fournier, M.: Training deep learning models with a multi-station approach and static aquifer attributes for groundwater level simulation: what is the best way to leverage regionalised information?, Hydrol. Earth Syst. Sci., 29, 841–861, https://doi.org/10.5194/hess-29-841-2025, 2025.

Collenteur, R. A., Haaf, E., Bakker, M., Liesch, T., Wunsch, A., Soonthornrangsan, J., White, J., Martin, N., Hugman, R., de Sousa, E., Vanden Berghe, D., Fan, X., Peterson, T. J., Bikše, J., Di Ciacca, A., Wang, X., Zheng, Y., Nölscher, M., Koch, J., Schneider, R., Benavides Höglund, N., Krishna Reddy Chidepudi, S., Henriot, A., Massei, N., Jardani, A., Rudolph, M. G., Rouhani, A., Gómez-Hernández, J. J., Jomaa, S., Pölz, A., Franken, T., Behbooei, M., Lin, J., and Meysami, R.: Data-driven modelling of hydraulic-head time series: results and lessons learned from the 2022 Groundwater Time Series Modelling Challenge, Hydrol. Earth Syst. Sci., 28, 5193–5208, https://doi.org/10.5194/hess-28-5193-2024, 2024.

Das, A., Kong, W., Leach, A., Mathur, S., Sen, R., and Yu, R.: Long-term Forecasting with TiDE: Time-series Dense Encoder, arXiv [preprint], https://doi.org/10.48550/arXiv.2304.08424, 2024.

European Environment Agency: European Digital Elevation Model (EU-DEM) v1.1, European Environment Agency (EEA) [code], 2018.

Gneiting, T. and Raftery, A. E.: Strictly proper scoring rules, prediction, and estimation, J. Am. Stat. Assoc., 102, 359–378, https://doi.org/10.1198/016214506000001437, 2007.

Gomez, M., Nölscher, M., Hartmann, A., and Broda, S.: Assessing groundwater level modelling using a 1-D convolutional neural network (CNN): linking model performances to geospatial and time series features, Hydrol. Earth Syst. Sci., 28, 4407–4425, https://doi.org/10.5194/hess-28-4407-2024, 2024.

Guzman, S. M., Paz, J. O., and Tagert, M. L. M.: The Use of NARX Neural Networks to Forecast Daily Groundwater Levels, Water Resour. Manag., 31, 1591–1603, https://doi.org/10.1007/s11269-017-1598-5, 2017.

Hermans, T., Goderniaux, P., Jougnot, D., Fleckenstein, J. H., Brunner, P., Nguyen, F., Linde, N., Huisman, J. A., Bour, O., Lopez Alvis, J., Hoffmann, R., Palacios, A., Cooke, A.-K., Pardo-Álvarez, Á., Blazevic, L., Pouladi, B., Haruzi, P., Fernandez Visentini, A., Nogueira, G. E. H., Tirado-Conde, J., Looms, M. C., Kenshilikova, M., Davy, P., and Le Borgne, T.: Advancing measurements and representations of subsurface heterogeneity and dynamic processes: towards 4D hydrogeology, Hydrol. Earth Syst. Sci., 27, 255–287, https://doi.org/10.5194/hess-27-255-2023, 2023.

Heudorfer, B., Liesch, T., and Broda, S.: On the challenges of global entity-aware deep learning models for groundwater level prediction, Hydrol. Earth Syst. Sci., 28, 525–543, https://doi.org/10.5194/hess-28-525-2024, 2024.

Izmailov, P., Podoprikhin, D., Garipov, T., Vetrov, D., and Wilson, A. G.: Averaging Weights Leads to Wider Optima and Better Generalization, arXiv [preprint], https://doi.org/10.48550/arXiv.1803.05407, 2019.

Jung, H., Saynisch-Wagner, J., and Schulz, S.: Can eXplainable AI Offer a New Perspective for Groundwater Recharge Estimation? – Global-Scale Modeling Using Neural Network, Water Resour. Res., 60, e2023WR036360, https://doi.org/10.1029/2023WR036360, 2024.

Kan, K., Aubet, F.-X., Januschowski, T., Park, Y., Benidis, K., Ruthotto, L., and Gasthaus, J.: Multivari-

ate Quantile Function Forecaster, arXiv [preprint], https://doi.org/10.48550/arXiv.2202.11316, 2022.

Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S., and Nearing, G.: Uncertainty estimation with deep learning for rainfall–runoff modeling, Hydrol. Earth Syst. Sci., 26, 1673–1693, https://doi.org/10.5194/hess-26-1673-2022, 2022.

Koya, S. R. and Roy, T.: Temporal Fusion Transformers for Streamflow Prediction: Value of Combining Attention with Recurrence, arXiv [preprint], https://doi.org/10.48550/arXiv.2305.12335, 2023.

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., and Nearing, G.: Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets, Hydrol. Earth Syst. Sci., 23, 5089–5110, https://doi.org/10.5194/hess-23-5089-2019, 2019.

Kratzert, F., Gauch, M., Klotz, D., and Nearing, G.: HESS Opinions: Never train a Long Short-Term Memory (LSTM) network on a single basin, Hydrol. Earth Syst. Sci., 28, 4187–4201, https://doi.org/10.5194/hess-28-4187-2024, 2024.

Kunz, S.: global-groundwater-models-main, GitHub [code], https://github.com/KunzstBGR/global-groundwater-models-main, last access: 14 July 2025.

Kværner, J. and Snilsberg, P.: Groundwater hydrology of boreal peatlands above a bedrock tunnel – Drainage impacts and surface water groundwater interactions, J. Hydrol., 403, 278–291, https://doi.org/10.1016/j.jhydrol.2011.04.006, 2011.

Li, X., Khandelwal, A., Jia, X., Cutler, K., Ghosh, R., Renganathan, A., Xu, S., Tayal, K., Nieber, J., Duffy, C., Steinbach, M., and Kumar, V.: Regionalization in a Global Hydrologic Deep Learning Model: From Physical Descriptors to Random Vectors, Water Resour. Res., 58, e2021WR031794, https://doi.org/10.1029/2021WR031794, 2022.

Lim, B., Arık, S. Ö., Loeff, N., and Pfister, T.: Temporal Fusion Transformers for interpretable multi-horizon time series forecasting, Int. J. Forecasting, 37, 1748–1764, https://doi.org/10.1016/j.ijforecast.2021.03.012, 2021.

Linardatos, P., Papastefanopoulos, V., Panagiotakopoulos, T., and Kotsiantis, S.: CO_2 concentration forecasting in smart cities using a hybrid ARIMA–TFT model on multivariate time series IoT data, Sci. Rep., 13, 17266, https://doi.org/10.1038/s41598-023-42346-0, 2023.

Linde, N., Ginsbourger, D., Irving, J., Nobile, F., and Doucet, A.: On uncertainty quantification in hydrogeology and hydrogeophysics, Adv. Water Resour., 110, 166–181, https://doi.org/10.1016/j.advwatres.2017.10.014, 2017.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V.: The M4 Competition: 100,000 time series and 61 forecasting methods, Int. J. Forecasting, 36, 54–74, https://doi.org/10.1016/j.ijforecast.2019.04.014, 2020.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V.: M5 accuracy competition: Results, findings, and conclusions, Int. J. Forecasting, 38, 1346–1364, https://doi.org/10.1016/j.ijforecast.2021.11.013, 2022.

Manzoni, A., Porta, G. M., Guadagnini, L., Guadagnini, A., and Riva, M.: Probabilistic reconstruction via machine-learning of the Po watershed aquifer system (Italy), Hydrogeol. J., 31, 1547–1563, https://doi.org/10.1007/s10040-023-02677-8, 2023.

Manzoni, A., Porta, G. M., Guadagnini, L., Guadagnini, A., and Riva, M.: A comprehensive framework for stochastic calibration and sensitivity analysis of large-scale groundwater models, Hydrol. Earth Syst. Sci., 28, 2661–2682, https://doi.org/10.5194/hess-28-2661-2024, 2024.

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., and Gupta, H. V.: What Role Does Hydrological Science Play in the Age of Machine Learning?, Water Resour. Res., 57, e2020WR028091, https://doi.org/10.1029/2020WR028091, 2021.

Nölscher, M., Mutz, M., and Broda, S.: Multiorder hydrologic Position for Europe – a Set of Features for Machine Learning and Analysis in Hydrology, Sci. Data, 9, 662, https://doi.org/10.1038/s41597-022-01787-4, 2022.

Olivares, K. G., Challu, C., Marcjasz, G., Weron, R., and Dubrawski, A.: Neural basis expansion analysis with exogenous variables: Forecasting electricity prices with NBEATSx, Int. J. Forecasting, 39, 884–900, https://doi.org/10.1016/j.ijforecast.2022.03.001, 2022.

Oosterwoud, M., van der Ploeg, M., van der Schaaf, S., and van der Zee, S.: Variation in hydrologic connectivity as a result of microtopography explained by discharge to catchment size relationship, Hydrol. Process., 31, 2683–2699, https://doi.org/10.1002/hyp.11164, 2017.

Oreshkin, B. N., Carpov, D., Chapados, N., and Bengio, Y.: N-BEATS: Neural basis expansion analysis for interpretable time series forecasting, arXiv [preprint], https://doi.org/10.48550/arXiv.1905.10437, 2020.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S.: PyTorch: An Imperative Style, High-Performance Deep Learning Library, arXiv [preprint], https://doi.org/10.48550/arXiv.1912.01703, 2019.

Pistocchi, A.: Leaf Area Index (MAPPE Model), European Commission, Joint Research Centre (JRC) PID [data set], https://cidportal.jrc.ec.europa.eu/ftp/jrc-opendata/MAPPE/MAPPE_Europe/LATEST/Vegetation/D_18_LAI/D_18_LAI.zip (last access: 3 February 2024), 2015.

Rangapuram, S. S., Seeger, M. W., Gasthaus, J., Stella, L., Wang, Y., and Januschowski, T.: Deep State Space Models for Time Series Forecasting. Advances in Neural Information Processing Systems, 31, https://proceedings.neurips.cc/paper/2018/file/5cf68969fb67aa6082363a6d4e6468e2-Paper.pdf (last access: 14 October 2024), 2018.

Razafimaharo, C., Krähenmann, S., Höpp, S., Rauthe, M., and Deutschländer, T.: New high-resolution gridded dataset of daily mean, minimum, and maximum temperature and relative humidity for Central Europe (HYRAS), Theor. Appl. Climatol., 142, 1531–1553, https://doi.org/10.1007/s00704-020-03388-w, 2020.

Salinas, D., Flunkert, V., Gasthaus, J., and Januschowski, T.: DeepAR: Probabilistic forecasting with autoregressive recurrent networks, Int. J. Forecasting, 36, 1181–1191, https://doi.org/10.1016/j.ijforecast.2019.07.001, 2020.

Schroeter, L. and Gläßer, C.: Analyses and monitoring of lignite mining lakes in Eastern Germany with spectral signatures of Landsat TM satellite data, Int. J. Coal Geol., 86, 27–39, https://doi.org/10.1016/j.coal.2011.01.005, 2011.

Sen, R., Yu, H.-F., and Dhillon, I. S.: Think Globally, Act Locally: A Deep Neural Network Approach to High-Dimensional Time Series Forecasting, arXiv [preprint], https://doi.org/10.48550/arXiv.1905.03806, 2019.

Tao, H., Hameed, M. M., Marhoon, H. A., Zounemat-Kermani, M., Heddam, S., Kim, S., Sulaiman, S. O., Tan, M. L., Sa'adi, Z., Mehr, A. D., Allawi, M. F., Abba, S., Zain, J. M., Falah, M. W., Jamei, M., Bokde, N. D., Bayatvarkeshi, M., Al-Mukhtar, M., Bhagat, S. K., Tiyasha, T., Khedher, K. M., Al-Ansari, N., Shahid, S., and Yaseen, Z. M.: Groundwater level prediction using machine learning models: A comprehensive review, Neurocomputing, 489, 271–308, https://doi.org/10.1016/j.neucom.2022.03.014, 2022.

Wright, L. and Demeure, N.: Ranger21: A synergistic deep learning optimizer, arXiv [preprint], https://doi.org/10.48550/arXiv.2106.13731, 2021.

Wu, H., Zhou, H., Long, M., and Wang, J.: Interpretable weather forecasting for worldwide stations with a unified deep model, Nature Machine Intelligence, 5, 602–611, https://doi.org/10.1038/s42256-023-00667-9, 2023.

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

Wunsch, A., Liesch, T., and Broda, S.: 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), Hydrol. Earth Syst. Sci., 25, 1671–1687, https://doi.org/10.5194/hess-25-1671-2021, 2021.

Wunsch, A., Liesch, T., and Broda, S.: Deep learning shows declining groundwater levels in Germany until 2100 due to climate change, Nat. Commun., 13, 1–13, https://doi.org/10.1038/s41467-022-28770-2, 2022a.

Wunsch, A., Liesch, T., and Broda, S.: Feature-based Groundwater Hydrograph Clustering Using Unsupervised Self-Organizing Map-Ensembles, Water Resour. Manag., 36, 39–54, https://doi.org/10.1007/s11269-021-03006-y, 2022b.