On the challenges of global entity-aware deep learning models for groundwater level prediction

Benedikt Heudorfer[1], Tanja Liesch[1], and Stefan Broda[2]

[1]Karlsruhe Institute of Technology (KIT), Institute of Applied Geosciences, Kaiserstr. 12, 76131 Karlsruhe, Germany
[2]Federal Institute for Geosciences and Natural Resources (BGR), Wilhelmstr. 25–30, 13593 Berlin, Germany

Correspondence: Benedikt Heudorfer (benedikt.heudorfer@kit.edu)

Abstract. The application of machine learning (ML) including deep learning models in hydrogeology to model and predict groundwater level in monitoring wells has gained some traction in recent years. ~~By~~ Right now, the dominant model class is so called single-well models, where one model is trained for each well separately. However, recent developments in neighbouring disciplines including hydrology (rainfall-runoff-modelling) have shown that global models, being able to incorporate data of several wells, may have advantages. These models are often called entity-aware models, as they usually rely on static data to differentiate the entities, i.e. groundwater wells in hydrogeology or catchments in surface hydrology. We test two kinds of static information to characterize the groundwater wells in a global, entity-aware deep learning model setup, first, environmental features that are continuously available and thus theoretically allow spatial generalization (regionalization), and second, time-series features that are derived from the past time series at the respective well. Moreover, we test random integer features as entity information for comparison. We use a published dataset of 108 groundwater wells in Germany, and evaluate the models' performances in terms of Nash-Sutcliffe efficiency (NSE) in an in-sample and an out-of-sample setting, representing temporal and spatial generalization. Our results show ~~,~~ that entity-aware models work well with a mean performance of NSE > 0.8 in an in-sample setting, thus being comparable to, or even outperforming single-well models. However, they do not generalize well spatially in an out-of-sample setting (mean NSE < 0.7, i.e. lower than a global model without entity information). ~~The reason for this potentially lies in the~~ Strikingly, all model variants, regardless of the type of static features used, perform basically equally well both in- and out-of-sample. The conclusion is, that the model in fact does not show entity awareness, but uses static features merely as unique identifiers, opening the research question regarding how to properly establish entity awareness in deep learning models. Potential future avenues lie in bigger datasets, as the relatively small number of wells in the dataset ~~, which~~ might not be enough to take full advantage of global models. ~~However, also~~ Also more research is needed to find meaningful static features for ML in hydrogeology.

1 Introduction

Groundwater is the primary drinking water resource in Germany (Hoelting and Coldewey, 2013) and a major one worldwide (WWAP, 2015). As such, it is under growing pressure due to e.g. climate change, increased drought frequencies or irrigation (WWAP, 2015). All of these transient drivers of change point to the ~~necessity~~ utmost importance of functional groundwater

25 management, for which groundwater level prediction models are a key tool . Currently, (Wada et al., 2010; Famiglietti, 2014; Bierkens and ~~V~~ . For functional national groundwater management, groundwater models need to be available on large or national scale. While some *conceptual* modelling approaches do exist that meet this criterium (Ahamed et al., 2022), they generally lack sufficient local-scale accuracy (Gleeson et al., 2021). Currently, distributed *numerical* models are the dominant hydrogeologic model class – distributed numerical models – are in groundwater level modelling. However, while being an excellent tool to answer

30 tangible questions in complex ~~hydrogeologic settings. But~~ local- to regional hydrogeologic settings, numerical models do not scale well. It is not trivial to parameterize larger scale numerical models because it is expensive. The process needs some degree of subjective sophistication and depends on amounts of unstructured geological or spatial data.

The ability of machine learning algorithms, especially neural networks, to incorporate large amounts of data and letting them sort out the complexity themselves in order to make highly accurate predictions on the point scale makes them a useful tool to

35 close this gap. With large-scale applicability also comes the potential to realize Prediction in Ungauged Aquifers (PUA), as put forward by Heudorfer et al. (2019) and Haaf et al. (2020, 2023), in a global, generalizable way. The idea of PUA is inspired by the widely known concept of Prediction in Ungauged Basins (PUB) in rainfall-runoff hydrology (Sivapalan et al., 2003; Hrachowitz et al., 2013). As the name tells, it refers to the aim of setting up models that are capable to predict groundwater levels in areas where no groundwater data is sampled, i.e. ungauged areas. This can be large areas with no or sparse monitoring, but

40 also areas in between groundwater monitoring wells in well-monitored regions. As groundwater dynamics can change within rather short distances due to heterogenous aquifer or infiltration conditions, such models could be used to create spatially continuous data out of point measurements at monitoring sites. Continuous groundwater level data are of great importance, as they serve as a basis for important decision-making tasks, such as deriving protection zones.

Even though the application of machine learning approaches in hydrogeology gained some traction in recent years, the field

45 is still emerging. In a recent review, Tao et al. (2022) recapitulated that neural network architectures often prove to be superior to other machine learning model classes. Thereby, state of the art groundwater level prediction with neural networks is mostly based on single-well models, where a single neural network model is trained for each groundwater well (Tao et al., 2022; Wunsch et al., 2021b, 2022a; Rajaee et al., 2019). While excellent fits can be achieved like this on the level of individual wells, the big drawback is that it is not possible to generalize or even regionalize with these models. A model that is capable

50 to overcome this drawback is called a "global model", where "global" means encompassing the whole dataset available. This model class made a first appearance in hydrogeology's sister field, hydrology, by the works of Kratzert et al. (2018, 2019a, b) on the task of rainfall-runoff modelling. They showed (Kratzert et al., 2019b) that a neural network can use static features (like time series features or environmental features, see below) to distinguish individual dynamic input features (meteorological) and target features (there: runoff) relations in a meaningful way, thereby allowing the model to generalize to locations with similar

55 combinations of static features. They called this setup an "entity-aware model" (Kratzert et al., 2019b), a term which is also common among various disciplines that use entity characteristics to model personalized prediction of responses for individual entities caused by external drivers (Ghosh et al., 2023).

While there have been approaches of global models for groundwater level modelling (e.g. Clark et al., 2022) or groundwater abstraction modelling (e.g. Majumdar et al., 2020, 2022), to the best of the authors' knowledge, the concept of entity-aware

global modelling has not been transferred to hydrogeology up to date. The reason might be that providing the model with a set of static features that is able to capture the hydrogeologic dynamics of the system in a conceptual way is much more difficult than in other disciplines, due to the complex nature of underground flow. Even though hydrology and hydrogeology are strongly related disciplines (Barthel, 2014), the question is whether the global entity-aware approach will work similarly well in hydrogeological conditions, which exhibit substantially more heterogeneous and "local" groundwater dynamics (Heudorfer et al., 2019), which is why nearby wells do not necessarily have high time series similarity (Wunsch et al., 2022b). It is not trivial to expressively link environmental static features taken from geospatial or geologic data products to groundwater dynamics (Haaf et al., 2020), also because of uncertainties in these products, which are often also only regionalized data from point measurements, such as for hydraulic conductivity or depth to groundwater.

For this reason, we test two different versions of a hydrogeologic global entity-aware model: the time series feature driven model (TSfeat model) and the environmental feature driven model (ENVfeat model). Both models have different potential applications. The ENVfeat model uses static environmental features that are available spatially and continuously (see section 2.2 for more information). It represents the gold standard of a fully generalizable and regionalizable model that we seek to achieve, in order to reach the overarching goal of PUA. However, geospatial data availability of sufficient quality to be used in machine learning is not always available. Also, above-mentioned lack of representativity of geospatial proxy data with regard to groundwater dynamics (Haaf et al., 2020) hypothetically hampers predictability in a hydrogeologic ENVfeat model. Thus, the TSfeat model is the viable alternative that retains the property of generalizability. The TSfeat model differentiates groundwater time series in multi-well prediction based on static time series features that are derived from the past groundwater time series itself (Heudorfer et al., 2019; Wunsch et al., 2022b, , see section 2.2). A TSfeat model can therefore hypothetically work on any existing monitoring network, being able to incorporate additional unseen groundwater data and thus generalize on groundwater data alone, despite lacking the ability to regionalize in the sense that it cannot predict groundwater head based on secondary data sources (i.e. geospatial data) like the ENVfeat model. We argue that time series features are best suited to describe the dynamics of an individual time series, and to best distinguish different groundwater dynamics from each other, thus hypothesizing that the TSfeat model outperforms the ENVfeat model.

In general, both "global" model types have the theoretical advantage of more data compared to single-well models, i.e. data of similar wells regarding dynamics can contribute to the training of e.g. wells with few training data and thus enhance their performance. To further test our hypotheses, we compare additional model setups, one where the set of environmental or time serie static features are replaced by a set of completely random static features (RNDfeat models), and one with no static features at all, only relying on the dynamic inputs (DYNonly model).

Regarding architecture, ~~we use~~ Long Short-Term Memory ~~neural networks (LSTM, Hochreiter and Schmidhuber, 1997) in the global model setup to learn the input-output relationship of meteorological forcing data and groundwater level, and combine it with a Multi-Layer Perceptron (MLP) for processing the static features. This is despite Wunsch et al. (2021b, 2022a) resorting to the Convolutional Neural Network~~ Neural Networks (LSTM) and Convolutional Neural Networks (CNN) ~~architecture for the single well setup because of better performance and stability. The reason is that LSTM showed overall better performance in preliminary experiments to this study (see figure A1)~~ are currently equally dominant model classes in ML-guided groundwater

level modelling, as they consistently deliver high-class performance while maintaining some degree of model simplicity (Tao et al., 2022). LSTMs are an especially common architecture for sequence modelling in hydrological settings (Kratzert et al., 2018, 201 , but also increasingly in groundwater (Tao et al., 2022; Rajaee et al., 2019). LSTMs are an improved version of simple Recurrent Neural Networks (RNN) and overcome RNNs drawback of limited memory of only few time steps (Bengio et al., 1994) which makes LSTMs suitable for time series modelling. It does so by introducing a cell memory state as well as an input gate, forget gate and output gate to control information storage and dissipation when flowing through the LSTM layer during training (Hochreiter and Schmidhuber, 1997). ~~LSTMs are a common architecture for sequence modelling, e.g. in hydrological settings (Kratzert et al., 2018, 2019a, b), and increasingly in groundwater (Tao et al., 2022; Rajaee et al., 2019).~~ For a while, Transformer architectures seemed promising candidates to generally supersede LSTMs in sequence modelling, but their suitability as a general-purpose method beyond its original domain (language modelling) are increasingly called into question because they can be outperformed by more simple linear neural models (Zeng et al., 2023), notably across the board (Das et al., 2023). In the neighbouring discipline of groundwater quality modelling, Extreme Gradient Boosting (XGB) prooved to be a powerful alternative (Bedi et al., 2020; Ransom et al., 2022; Haggerty et al., 2023). However, groundwater *quality* modelling is a related but different field from groundwater *level* modeling with a number of significant differences, making methods not directly transferable. Thus, for the moment, LSTM and CNN methods remain the state of the art in groundwater level modelling. In the present study, we use a LSTM in the global model setup to learn the input-output relationship of meteorological forcing data and groundwater level, and combine it with a Multi-Layer Perceptron (MLP) for processing the static features. This is despite precursor studies Wunsch et al. (2021b, 2022a) resorted to the CNN architecture for the single well setup, as in their study, better performance and stability was achieved with CNNs. However, in the (global) use case presented here, LSTM showed overall better performance in preliminary experiments (see figure A1), thus a LSTM architecture was chosen.

The aim of this study is to test whether the concept of entity-aware global deep learning modeling is transferable to groundwater level prediction, and, if so, which set of static features are best suited to do so. First, we introduce the dynamic (section 2.1) and static (section 2.2) feature data used in the study. We then give an overview of the model architecture and optimization strategy (section 3.1) and an outline of the experimental design (section 3.3), followed by a brief ~~introduce~~ introduction of the state-of-the-art learning rate scheduling method to help learning (section 3.2). We then compare model performance of the ENVfeat, TSfeat, RNDfeat and DYNonly model variants (section 4.1) and discuss possible reasons. We further analyse feature importance (section 4.2) to get to the bottom of performance differences. The paper ends with concluding remarks and an outlook in section 5.

2 Data

2.1 Dynamic feature data and study area

Regarding groundwater level data, the dataset of Wunsch et al. (2022a) is used in this study. This dataset consists of 118 weekly groundwater level time series from the uppermost ~~,~~ unconfined aquifer layer, ~~which~~ whose groundwater dynamics are mainly

dominated by climate forcing. The wells are relatively evenly distributed across Germany~~and were picked to make the spatial coverage as representative as possible within the national borders. Also, only wells that proved high forecasting accuracy in~~

130 ~~past experiments were used in this dataset. In addition, a low percentage of gaps was filled with appropriate interpolation tools~~, as can be seen in figure A2. The dataset was primarily chosen because it is pre-published and readily available from an open source repository, enabling reproducibility and ease of publication. For additional information on the dataset and details on data preprocessing routines please refer to Wunsch et al. (2022a). Based on this readily available dataset, the only additional preprocessing step was to exclude time series with start dates after 01.01.2000 and end dates before 31.12.2015. This was done

135 to make sure that enough data is present for good model results. Also, four individual time series were excluded manually because of missing environmental static feature data (see section 2.2). This resulted in a total number of 108 groundwater time series used in this study.

Dynamic input features used in this study are precipitation (P), temperature (T) and relative humidity (rH) from HYRAS 3.0 (Rauthe et al., 2013; Frick et al., 2014), which proved its suitability in several previous studies (Wunsch et al., 2021b, 2022a)

140 , as well as an annual sinusoidal curve fitted to temperature (Tsin), which ~~proved~~ was also shown to be a valuable driving variable (Wunsch et al., 2021b). HYRAS 3.0 is a 1x1 km gridded meteorological dataset covering German national territory, with data ranging from 1951 to 2015. The dataset is essentially the same as in Wunsch et al. (2022a), however HYRAS' relative humidity and the fitted sinusoidal curve was used here additionally.

2.2 Static feature data

145 It is well known that no single static feature is able to describe the totality of control on groundwater dynamics, but a combination of features can provide a good approximation (Heudorfer et al., 2019). Consequently, exhaustive yet compact sets of static input features were compiled. Thereby, two separate sets of static features were fed into two separate model setups: time series features and environmental features. The first type (time series features) are quantitative metrics calculated from the groundwater time series themselves and express certain aspects of dynamics in these time series. There is a long history

150 of studies in hydrology and to a somewhat lesser extent in hydrogeology (see introduction of (Heudorfer et al., 2019) for a brief review) dedicated to finding, improving and analyzing which time series features best depict certain aspects of time series dynamics, or the totality of time series dynamics in a reduced set of time series features. Oftentimes redundancy analyses, correlation analyses, dimensionality reduction or similar methods are conducted to determine a suitable set of features. As a conceptual decision, we use the set of time series features devised by Wunsch et al. (2022b). This set constitutes a small and

155 manageable, redundancy-reduced set of time series features that was furthermore successfully applied in past studies to cluster time series in the larger dataset of wells (Wunsch and Liesch, 2020; Wunsch et al., 2022b) from which the sample of wells used in this study are taken (see section 2.1). The full list of time series features used in this study, along with descriptions, can be found in table 1.

The second type of features, environmental features, are descriptors of the hydrogeological, physiographic and climatic

160 functioning of the underground and landscape. They are proxies for environmental factors controlling groundwater recharge and flow and thus the dynamics in groundwater time series (Haaf et al., 2020). To be able to reach the stated goal of PUA

Table 1. Static time series features used in the TSfeat model variant. Features were derived from the groundwater level (see section 2.1) from the beginning of the time series until 2011 (to exclude the test period from 2012 onwards). Table taken from Wunsch et al. (2022b).

short name	feature name	description	citation
RR	Range Ratio	Detection of superimposed long-periodic signals, also sensitive to outliers, calculated as the ratio of the mean annual range to the overall range	Wunsch et al. (2022b)
Skew	Skewness	Boundedness, inhomogeneities, outliers, asymmetry of the probability distribution	Wunsch et al. (2022b)
P52	Annual Periodicity	Strength of the annual cycle, calculated by correlating (Pearson) the mean annual (52 weeks) periodicity with the complete time series	Wunsch et al. (2022b)
SDdiff	SDdiff	Flashiness, frequency, and rapidity of short-term changes, calculated as the standard deviation of all first derivatives	Wunsch et al. (2022b)
LRec	Longest Recession	(unnaturally) long descending heads, longest sequence without rising head values	Wunsch et al. (2022b)
jumps	Jumps	Inhomogeneities/breaks, partly also variability, calculated as the absolute and standardized maximum change of the mean of two successive years	Wunsch et al. (2022b)
SB	Seasonal Behaviour	Position of the maximum in the annual cycle, agreement with the expected average seasonality (Min in September, Max in March)	Wunsch et al. (2022b)
med01	Median[0,1]	Boundedness, median after scaling to [0,1]	Heudorfer et al. (2019)
HPD	High Pulse Duration	Average duration of heads exceeding the 80th percentile of non-exceedance	Richter et al. (1996)

(see section 1), it is important that environmental features used in the model are spatially continuously available across the study domain (Germany). Thus, only nationwide geospatial datasets are considered in the selection ~~which~~ that are, for the sake of reproducibility, freely available. Moreover, we use datasets that are not too fine-grained (in the case of categorical data), in order to ensure that each category is represented by one or more monitoring wells. This means that when multiple source datasets for the same type of environmental feature are available, the one with the better (usually higher) degree of aggregation was chosen. For example, for soil type, there is a product called "buek200" with more than 550 categories (BGR, 2018), and another product called "buek5000" (BGR, 2005) which is a generalized version of buek200 and has only 23 categories. In that instance, Buek5000 was chosen because communality of soil type categories between groundwater locations would be impossible with buek200, given the limited size of our groundwater data set.

Table 2. Environmental static features used in the ENVfeat model variant. Hydrogeologic, soil, topographic and land cover features as well as ~~ETPpot~~ ETpot were derived from map products by sampling the map value at the location of the groundwater well. Climatic static features with exception of ~~ETPpot~~ ETpot were derived from the meteorologic dynamic input features (see section 2.1) in the period from 1991 (to match ~~ETPpot~~ ETpot data availability) until 2011 (to exclude the test period from 2012 onwards).

Type	Short name	Description	Unit	Citation
Hydrogeologic	Recharge	Mean annual groundwater recharge rates 1961-1990	mm	BGR (2019)
	Percolation	Mean annual groundwater percolation rates 1961-1990	mm	BGR (2003)
	Hygeo_division	~~Divisions/areas of~~ Division of major hydrogeologic units, defined by similar hydrogeologic properties, groundwater conditions and geologic genesis	categ.	BGR & SGD (2015)
	Conductivity	~~Hydraulic conductivity~~ Predominant hydraulic conductivity rates (kf value) of the aquifer	m/s	BGR & SGD (2019)
	Aquifer type	~~Classes~~ Classification of aquifer types (4 categories: porous, karstic, fractured, mixed)	categ.	BGR & SGD (2019)
Soil	Soil type	~~Soil type~~ Classification of predominant soil types (23 categories)	categ.	BGR (2018) ~~-~~
	Clay	Clay content of the soil in weight fraction	%	Hengl et al. (2017)
	Sand	Sand content of the soil in weight fraction	%	Hengl et al. (2017)
Topographic/ drainage	TWI	Topographic Wetness Index; estimate where water will accumulate in area with elevation differences	—	calculated after Beven and Kirkby (1979)
	Divide to stream	Distance from hypothetical groundwater catchment divide to nearest stream (hydrologic order 3) at the groundwater well location	m	Noelscher et al. (2022)
	Lateral position	Relative position of the groundwater well lateral along the divide-to-stream stretch (hydrologic order 3)	—	Noelscher et al. (2022)
	Stream distance	Distance from the groundwater well to nearest stream (hydrologic order 3)	m	Noelscher et al. (2022)
Land cover	CLC land cover	CORINE ~~land cover classes (5 ha scale~~ Land Cover classification (CLC, 14 categories)	categ.	EEA (2018)
Climatic	Tmean	Mean annual average temperature	°C	~~self-derived~~ calculated from T
	Psum	Mean annual sum of precipitation	mm	~~self-derived~~ calculated from P
	~~ETPpot~~ ETpot	Mean annual sum of potential evapotranspiration	mm	DWD Climate Data Center (CDC) (2023)
	rHmean	Mean annual average air humidity	%	

In surface hydrology, the science of identifying major controlling factors for river flow systems is mature enough to yield large-scale or even global selection datasets of environmental features with which river flow can be predicted and explained in an exhaustive way (e.g. Addor et al., 2017; Linke et al., 2019; Kratzert et al., 2019b, 2023). In hydrogeology, this is not the case yet. No comparable selection of tried and tested sets of environmental features controlling groundwater dynamics is available that could be used in machine learning based prediction of groundwater head. In light of that, we compiled a first selection that is summarized in table 2. It was assembled primarily by conceptual decision from available geospatial datasets to cover the five major domains of control, namely hydrogeology, soil, topographic drainage, land cover and climate. Specifically, climate can be of transient nature, which is why it was used as a long-term average over the whole study period. Also, land cover can technically be transient e.g. in highly dynamic developing countries. However, land cover in Germany can be assumed to be static within the study period with a high degree of confidence. Some factors such as e.g. depth to groundwater, which are important factors based on conceptional understanding, are omitted, due to above-mentioned data availability reasons. Thus, this selection should be seen as a starting point to serve the proof of concept in this study.

~~For testing our hypothesis~~To ground-truth the effect of the static environmental and time series features, a third set of static features was used, ~~namely features with~~ by replacing the number of static features with random counterparts of equal size, i.e. sets of randomly generated integers in the range of 0 to 9. This was done in two variants, ~~first,~~ one set with nine random integers to equal the number of time series features (RNDfeat9), and one set with 18 random integers to equal the number of environmental features (RNDfeat18). These numbers were chosen to make sure that all aspects but the values of the static features themselves match the model setup of the TSfeat and the ENVfeat models to exclude other influences.

All numeric static features (time series, environmental or random) were standard scaled before feeding them into the model. All categorical static features (only environmental) were one-hot-encoded.

3 Methods

3.1 Model architecture and optimization strategy

We use LSTMs in a global entity-aware model setup to learn the input-output relationship between dynamic (meteorological) input features and groundwater level as the target feature. LSTM was chosen above CNN because it proved to be superior to CNN in the global entity-aware model setup in preliminary studies (see figure A1 for a comparison). To allow the global model to distinguish between different groundwater dynamics of individual wells, static input features that differentiate the wells must be fused to the dynamic (meteorological) input features. Different approaches exist to accomplish this data fusion. Most notably, Kratzert et al. (2019a, b) provide two separate variants to accomplish data fusion. The first is the basic variant where static features are simply concatenated to the meteorological inputs at each time step, together entering the model through the same input layer. The second is more sophisticated with a modified LSTM layer where static features control the input gate, and the dynamic features control the forget gate, output gate and memory. Even though the modified LSTM layer variant provides desirable levels of interpretability, the basic ~~model~~ data fusion variant notably outperformed it (Kratzert et al., 2019b). Thus, we disregard the modified LSTM layer variant in this study. ~~However, instead of the oversimplistic design of Kratzert's basic~~ But

as discussed in (Miebs et al., 2020), also the basic data fusion variant is not an optimal choice for RNN architectures, since such an approach leads to a significant increase in the number of RNN parameters due to the fact that duplicated static features are evaluated each time for every sequence, while not adding any meaningful additional information. As a consequence, training becomes more memory- and time-consuming at comparable outcome (Miebs et al., 2020). Thus, instead of the simplistic basic duplication variant, we ~~present~~ use a model architecture where data fusion of dynamic and static input features is achieved by providing separate model threads that process the dynamic and static inputs individually and are later concatenated (fig. 1)~~. For~~, as also used and put forward by e.g. Miebs et al. (2020); Liu et al. (2022).

In this model, for every time step being processed~~by the model~~, a sequence of the previous 52 time steps (making up one full year) of the four dynamic input features P, T, Tsin and rH is given to a LSTM layer of size 128 in the dynamic model thread. In the same time step, one set of static feature values, associated with the well whose groundwater head is currently being processed, is fed into a Multi-Layer Perceptron (MLP) ~~, with~~ of one fully connected (Dense) layer of size 128 in the static model thread. Subsequently, outputs of both threads are concatenated and fed into another MLP with a Dense layer of size 256, which again feeds into an output Dense layer of size one. In the whole architecture, all neural layers (despite the output) are followed by a dropout layer with a dropout rate of 0.1 for regularization.
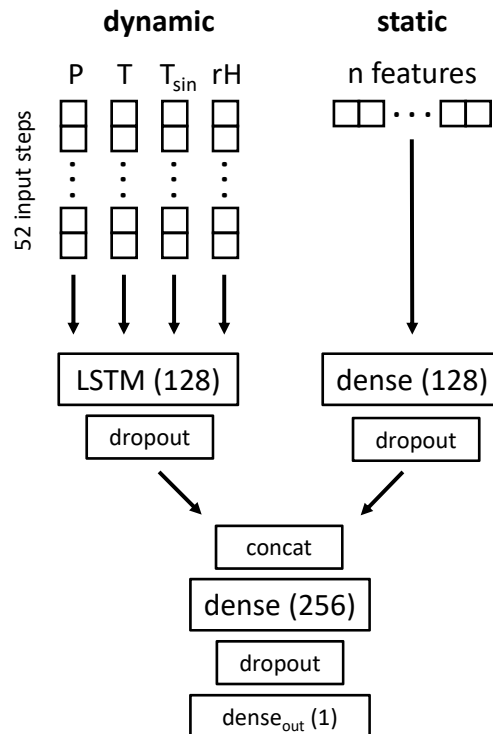
Figure 1. The double threaded global entity-aware model architecture introduced in this paper. Hyperparameters not found in the figure are reported in the text or can be found in the associated code.

For every well separately, both groundwater and meteorological time series were split into three parts: Training set, validation set and test set. To ensure comparability of performance between wells in light of interannual (or even interdecadal) fluctuations in groundwater, we chose to set a fixed time period for the validation and test sets. The validation period was scheduled 01.01.2008 - 31.12.2011, and the test period was scheduled 01.01.2012 - 31.12.2015, which is four years each. The training period was left open towards the past, meaning the model takes various time series lengths during training, under the sole condition that training data is available at least from 2000 onwards, i.e. at least eight years of training data (see section 2.1).

The model was optimized with the Adam optimizer (Kingma and Ba, 2014) on the Mean Squared Error (MSE) loss function during training, and later evaluated based on Nash-Sutcliffe Efficiency (NSE) calculated from model errors in the test set, while the Coefficient of Determination (R^2) as well as the Root Mean Squared Error (RMSE) are reported in this paper merely for comparison. This way of calculating the NSE metric is directly adapted from (Wunsch et al., 2022a) to ensure comparability with the NSE values reported therein for a single-station model setup. However, for the global model setup, specific considerations have to be taken into account. As Kratzert et al. (2019b) rightfully noted, MSE and NSE are both squared error loss functions, with the difference that the NSE is normalized by variance. This implies that MSE and NSE are linearly correlated and will yield the same model results in a single well model setup. However, in multi-well model setups this linear relationship is lost because of differing mean and variance of observed groundwater levels in different wells, heavily altering un-normalized MSE scores. As a remedy, Kratzert et al. (2019b) introduces a custom, basin-averaged NSE, where during training NSE loss is calculated per basis and averaged afterwards. We applied a simpler solution with the same effect, by standardizing each groundwater time series separately int z-scores during preprocessing. This way, the linear relationship between MSE and NSE is restored in the multi-well model setup when using MSE as a loss function during training and NSE as an evaluation metric, while at the same time avoiding the use of computationally expensive custom loss functions.

3.2 Learning rate scheduling

To avoid rapid overfitting and exploding gradients, a behavior not uncommon in LSTM models (e.g. Goodfellow et al. (2016)), we used a relatively large batch size of 512 (3 permille of the 160,415 samples in the training set) to make the learning process less stochastic and thus more stable. Moreover, we decreased the overall learning rate to 0.0003 (from the Keras default of 0.001). To further improve learning efficiency, we applied a learning rate schedule (LRS) combining a learning rate warmup with subsequent learning rate decay. Warmup is a limited phase in the beginning of training where the learning rate is gradually increased until it reaches a target learning rate (lr_{target}). This fights early overfitting by reducing the primacy effect of the first training examples learned by the model, since in unbiased datasets, the model can learn "superstitions" from the first learning examples otherwise uncommon in the dataset. We use a warmup period of 1 epoch, starting from $lr_0 = 0$ to the above-mentioned lr_{target}=0.0003. After the initially high lr_{target} is reached, it is slowly reduced again. This strategy - warmup periods followed by initial high learning rates - was shown to improve the performance of neural networks (Smith and Topin, 2018; Li et al., 2019). In our case, it led to a strong stabilization of the loss curve (see fig. 2) at unchanged performance. We used a cosine shaped learning rate decay after warmup using the formula below, slightly differing from ready-to-use implementations e.g. in tensorflow.

$$lr_i = 0.5 * lr_{target}(1 + cos(\frac{\pi * (i_i + i_{warmup})}{i_{total} - warmup}))$$ (1)

Where lr_i is the learning rate at the current batch step, lr_{target} is 0.0003, i_i is the integer of the current batch step, i_{warmup} is the number of batch steps during warmup (number of total training samples divided by batch size) and i_{total} is the total number of batch steps during all training epochs. In our case, training spanned 30 epochs. The shape of the LRS is illustrated in figure A3.
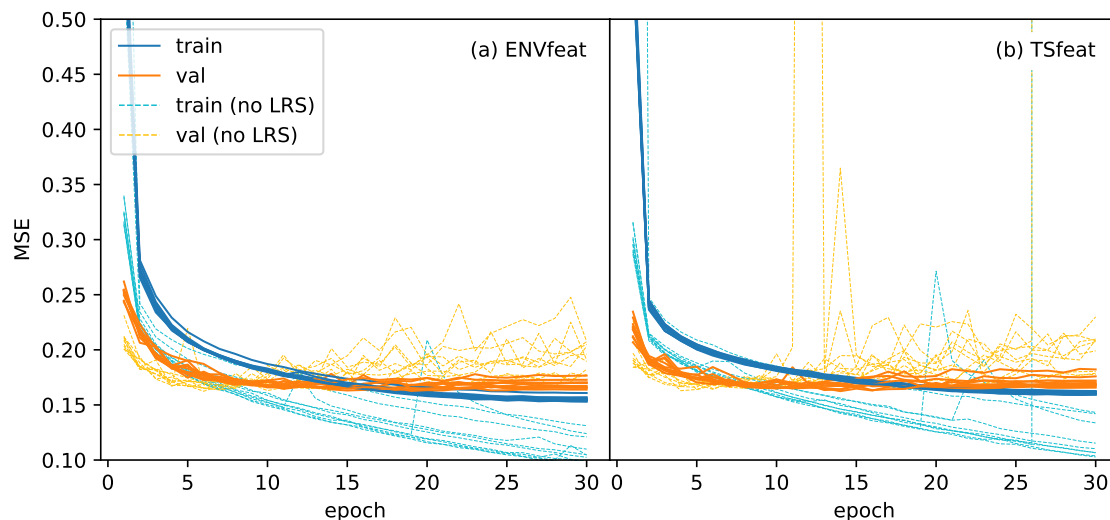
Figure 2. Losscurves of the training and validation period for the 10 different seed initialization runs for the (a) ENVfeat model and (b) TSfeat model, each with and without the learning rate scheduling described in section 3.1. The figure shows a strong stabilization effect on losscurves.

Figure 2 shows the loss curves for 10 seeds with and without the LRS for the in-sample ENVfeat and TSfeat models (see figure A4 for loss curves of the other model variants). The TSfeat model does have a slight MSE improvement with a mean of 0.002 when using LRS, but the ENVfeat model actually performs slightly worse with a mean MSE decrease of 0.008 when using LRS. The same can be said when considering NSE in the test period, where mean performance decreases by 0.0005 with LRS. These changes in performance are negligibly small fluctuations around zero, and in essence, it can be said that model performance is basically the same with or without LRS. Thus, no greatly improved performance can be observed, as was achieved by e.g. Smith and Topin (2018) and Li et al. (2019). However, introducing LRS comes with a strong stabilization effect on the loss curve (~~figure 2).~~ fig. 2): While learning without LRS shows tendencies of rapid overfitting as well as ~~to~~ of heavily exploding gradients, none of that is visible when using LRS. Loss curves with LRS are near-ideal, with train and validation loss curves approaching each other gently, and validation loss curve never significantly increasing after reaching

its minimum. The greatly increased stability of the loss curve implies much better generalization abilities of the model when applying LRS and shows the big advantage of this technique.

Figure 2 also shows some degree of bias in the validation dataset. This can be seen by the faster initial convergence of the validation loss curve. This is most likely due to the fact that validation data was not stratified, i.e. uniformly sampled over the whole time period or frequency spectrum. Instead, we used the fixed time period of 2008-2011 (and test period is 2012-2015, see section 3.1). These periods constitute the most recent years, and were consciously chosen as fixed because the ultimate aim is forward prediction of groundwater levels, where the most recent groundwater levels are most representative for future predictions (as opposed of choosing rolling time periods for validation and testing).

3.3 Experimental design

After evaluating the effect of the LRS approach on training, we set up three different global model variants with static features. Using the model architecture described in section 3.1, we either used static time series features, static environmental features or random static features in the static model thread to build the time series feature model (TSfeat model), the environmental feature model (ENVfeat model), and the random feature model (RNDfeat model) respectively. The RNDmodel was run with two different numbers of features, i.e. 9 to be consistent with the number of features in the TSfeat model, and 18 as used in the ENVmodel. Additionally, we run a ground truthed model variant without any static features, i.e. only the dynamic strand (DYNonly model) as described in section 3.1.

To test the models' performance, we first run all models in an in-sample (IS) setting where all wells are used for training and performance is evaluated for each well separately based on the NSE score in the test period. We then compare their test score (NSE) to the results of the single station models of Wunsch et al. (2022a), i.e. models trained and hyperparameter-optimized on every well separately. Importantly, we took their published results for this and did not rerun any of the single station models. Also, it has to be noted that comparing with the single station scores of Wunsch et al. (2022a) is not benchmarking in a narrow sense, since some differences beyond the model architecture exist, e.g. Wunsch et al. (2022a) did not use rH as an input (which is used here) and optimized the input sequence length (which is fixed to 52 here).

To further test desired capabilities of spatial generalization, all global models were additionally run in an out-of-sample (OOS) setting, where the models are tested on unseen data, i.e. wells not used for training. To be precise, for every well the models are trained leaving the wells' entire data out of training, and then predicting the wells' score using only the wells' test data. Practically, this would equate to applying a Leave-One-Out (LOO) cross validation, but due to computational constraints, we used 10-fold cross validation instead to test the described out-of-sample performance. To ensure robustness of the results, all models were run with 10 different seeds for random weight initialization on both settings (in-sample and out-of-sample).

Finally, to understand the inner workings of the model, and how its performance relates to its input data, Permutation Feature Importance (?Fisher et al., 2019) (PFI, Fisher et al., 2019) was applied. ThebreyPFI was chosen over alternative methods like SHAP (Lundberg and Lee, 2017) due to their high computational cost. With PFI, we measure the importance of features by successively taking every individual input feature (dynamic or static) of the trained model, permuting it by shuffling it randomly,

and then calculating the models' error (here: MSE) in the test data. A strong increase in the models' error equates to a high importance of the feature being shuffled. This was repeated for each of the 10 randomized initialization seeds.

An additional side experiment was to compare the LSTM-based model to a modified version where the LSTM layer is replaced by a CNN layer, as suggested by the results of (Wunsch et al., 2021b). However, the results of the CNN variants showed consistently lower performance in the used model setup (see fig. A1), so the results shown in the following focus on the ones achieved with the LSTM models.

4 Results and Discussion

4.1 Performance comparison of model variants

A side-by-side comparison of all global model scores and the single-well model scores of Wunsch et al. (2022a) can be seen in figure 3. The mean, lower (10%) and upper (90%)-percentile NSE of the 10 ensemble members for all model variants are shown in table 3.

In the in-sample validation (IS), all global models with the exception of the DYNonly model perform almost identically with only minor differences, at the level of statistical noise. Only the RNDfeat9 model seems to show a slightly lower performance. Two things are somewhat unexpected in this result, first, that the TSmodel does not outperform the ENVmodel, and second and even more striking, that the RNDfeat models can keep up with the performance of the models with "meaningful" static features. This result corresponds to the findings of ?Li et al. (2022), who replaced their static environmental features with random counterparts and found similar or even improved performance for rainfall-runoff modelling in CAMELS-basins. We speculate that the models use the static features solely as a kind of "unique identifier" for the wells, thus, it doesn't seem to matter if the static values represent some meaningful information (in terms of generalization) or not. This shows that our model is not able to learn from wells with similar static features, probably due to the number and choice of wells in the considered dataset. The reason for the inferior IS performance of the DYNonly model however seems obvious: since no static features are provided, the model is not able to distinguish between the different wells and rather fits to some average behaviour of all wells.

All global models with static features also slightly beat the scores of the single-well models. This result confirms observations by Kratzert et al. (2019a, b), who also observe better scores of global models over single-station models in rainfall-runoff prediction. This seemingly contradictory results – after all, single-well models are specifically optimized for the one specific location and should know this location best – is often attributed to the fact that, contrary to traditional hydrogeologic or hydrologic models, machine learning models benefit from additional data. However, with the RNDfeat models being as good as the TSfeat and ENVfeat models, we can widely exclude this as a reason in our case (maybe with the exception of the TSfeat model, which seems to perform slightly better than its RNDfeat9 counterpart). Benefiting from additional data or additional wells, respectively, would presuppose, that the model is able to identify wells that react similar to meteorological inputs based on the static features provided. With the random number features being different for all wells and no meaningful similarity depicted in them, this cannot be the case.

335

Figure 3. Cumulative distributions of NSE for the model variants ENVfeat, TSfeat, RNDfeat9, RNDfeat18 and DYNonly in in-sample mode (IS) and out-of-sample mode (OOS) against the performance of the single-well models by Wunsch et al. (2022a) (*). Lines represent sorted median NSE scores of 10 ensemble members. A version that includes the ensemble ranges as envelopes is shown in Fig. A5.

The differences to the single-well models might just as well be attributed to the different meteorological input parameters (rH and Tsin used additionally) and different optimization strategy. Moreover, despite being better on average, we observe a significant tailing in all global model setups, which the single-well models do not experience to the same degree (figure 3). The tail is the only part of the dataset where single-well models outperform both global models in 18 wells.

340   Figure 3 also shows the out-of-sample (OOS) performance of the global models, represented by 10-fold cross validation runs. As expected, there is a sharp decrease in model performance when a time series' test data is predicted by a model that never saw the time series' training data. However, with a mean NSE of 0.675 (down from 0.8213, see table 3) for the ENVfeat model, and a mean NSE of 0.6914 (down from 0.8229) for the TSfeat model, on average the OOS models perform surprisingly well, especially since some of the performance loss might be attributed to the compromise of 10-fold CV (instead

345 of Leave-One-Out-CV), where about 10% of possible training data is lost compared to the IS setting. These indications of model robustness are counteracted by the large seeding spread associated with both OOS models (fig. 4 or fig A5), and an amplification of the tailing effect. Also, as in the IS case, the RNDfeat models perform nearly identical to the ENVfeat and

Table 3. Mean, lower (10%) and upper (90%)-percentile NSE of the 10 ensemble members for all model variants as well as the mean NSE for single-well models as published in Wunsch et al. (2022a). R^2 and RMSE show similar patterns as NSE and are reported for comparison, but not discussed in-text.

variant	NSE (Q_{10})	NSE (Q_{50})	NSE (Q_{90})	R^2 (Q_{50})	RMSE (Q_{50})
single-well	-	**0.8134**	-	0.8255	0.2961
ENVfeat (IS)	0.8026	**0.8213**	0.8397	0.8418	0.2656
RNDfeat18 (IS)	0.7909	**0.8215**	0.8457	0.8354	0.2673
TSfeat (IS)	0.8028	**0.8229**	0.8395	0.8402	0.2677
RNDfeat9 (IS)	0.7777	**0.8135**	0.8399	0.8274	0.2746
DYNonly (IS)	0.7094	**0.7347**	0.7554	0.7670	0.3580
ENVfeat (OOS)	0.3977	**0.6750**	0.7685	0.5102	0.3437
RNDfeat18 (OOS)	0.4156	**0.6767**	0.7707	0.5145	0.3396
TSfeat (OOS)	0.4590	**0.6914**	0.7697	0.5491	0.3403
RNDfeat9 (OOS)	0.4433	**0.6817**	0.7710	0.5386	0.3401
DYNonly (OOS)	0.7103	**0.7326**	0.7518	0.7583	0.3574

the TSfeat model (fig. 3), indicating that the model is not able to truly generalize well based on the provided static features, neither on the environmental nor on the time-series features. ~~While the TSfeat model performs~~ The TSfeat model does perform
350   at least slightly better than its RNDfeat9 counterpart, which could at least partly support our initial hypothesis that the TSfeat model would outperform the ENVfeat model since static time series features are deemed to be informationally more complete and static environmental features suffer from high uncertainty~~,~~. But the differences are minor and ~~could also be influenced by~~ might be indistinguishable from noise due to the relatively low number of 10 ensemble members, showing a large range (fig. 4 or fig A5). In the median, the ranges of NSE values at individual wells for different model seed realizations for the ENVfeat
355   and TSfeat model are around 0.5 in the OOS setting, with minimum range values around 0.08 and maximum range values of more than 1. Even though the spread is one magnitude smaller in the IS setting for all models (medians hovering around 0.05, see figure 4), this is some significant spread and shows that even if different model runs have the same NSE on the global level, they will have significantly different outcomes on the level of the individual well.

     Most strikingly, the DYNonly model, having no static features at all, clearly outperforms all models with static features in
360   the OOS setting (fig. 3, tab. 3). This indicates that the static features even seem to hamper the global entity-aware models to learn a meaningful relationship that is generalizable from the static features in the out-of-sample setting. Furthermore, the OOS performance of the DYNonly model is equally good as its IS performance (down to the third decimal, see table 3), even though it relies on 10 % less training data. This implies information saturation, meaning that all information needed to reach the IS performance of the DYNonly model can be found in a significantly smaller subset of the dataset.
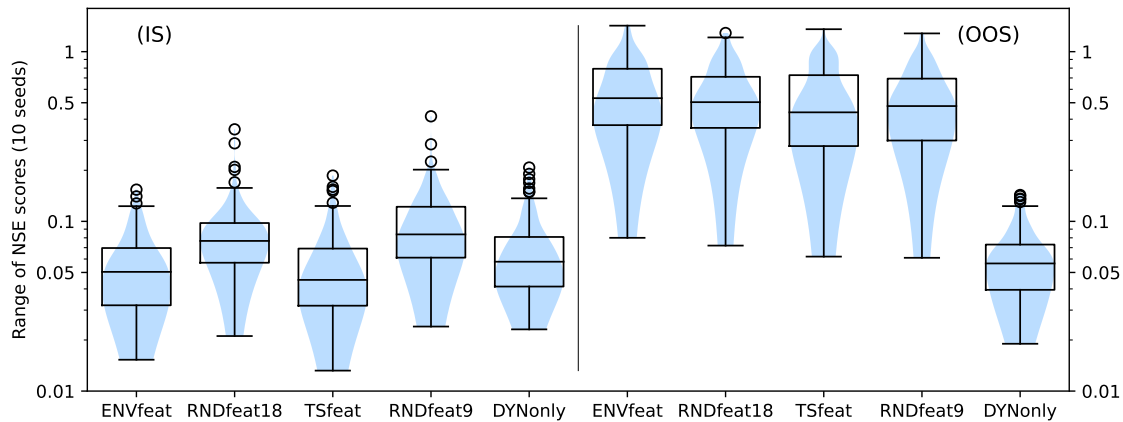
Figure 4. Range of NSE scores of the 10 ensemble members of all model variants in in-sample mode (IS) and out-of-sample mode (OOS).

365   More interesting insights can be drawn on the level of individual well predictions (fig. 5, see supplements for all other wells). Basically, there are two groups of wells, where all wells show more or less the behaviour of one of the two groups, with smooth transitions in between. In the first group, exemplarily shown by well BB_30400591 (fig. 5, top), the predictions in the IS setting of the ENVfeat, TSfeat, and RNDfeat models match rather well the observations with NSE for all models above 0.8, while the DYNonly model clearly fails (NSE = 0.169). Moreover, the predictions of the ENVfeat, TSfeat, and RNDfeat models are

370 quite similar, confirming their overall similar performance. Thus, the models seem to obtain important information from the static features, that allow to better train the model to the individual behaviour of the well, while (as already postulated above) the kind of static information, environmental, time-series or random, doesn't seem to matter. On the other hand, the DYNonly model obviously lacks information about the special behaviour of the well, and probably predicts some average reactions to the inputs, that do not work well for the considered well. In the OOS setting, the predictions of all models are quite similar,

375 or in other words, the predictions of the ENVfeat, TSfeat and RNDfeat approach the one of DYNonly model, which is nearly the same as in in-sample setting. The ENVfeat, TSfeat and RNDfeat have obviously lost their ability to predict the individual behaviour of the considered well, as the well was not included into training, and the model was obviously not able to generalize the relevant information from other wells' static data. Thus, there is a drastic drop in model performance between IS and OOS setting. Other good representatives for this group are e.g. well BW_107-666-2, and well SH_10L53126001 (see supplements).

380   In the second group, represented by well SH_10L62060004 (fig. 5, botton), all model predictions in the IS and OOS setting, including the DYNonly model, are quite similar. All models perform well with NSE > 0.8 and there is no obvious performance drop between IS ans OOS. Our interpretation is, that these are wells show a more "average" behaviour in terms of their reaction to the meteorological inputs, i.e. the "average" reaction to the meteorological inputs that is learned by the DYNonly model. Therefore, the additional information provided by the static parameters does not improve model performance in IS

385 setting. Conversely, it also doesn't negatively influence model performance when this information is missing OOS, explaining

the absent or minor drop in performance. Other good representatives for this group are e.g. well BW_124-068-9, and well NI_200001722.
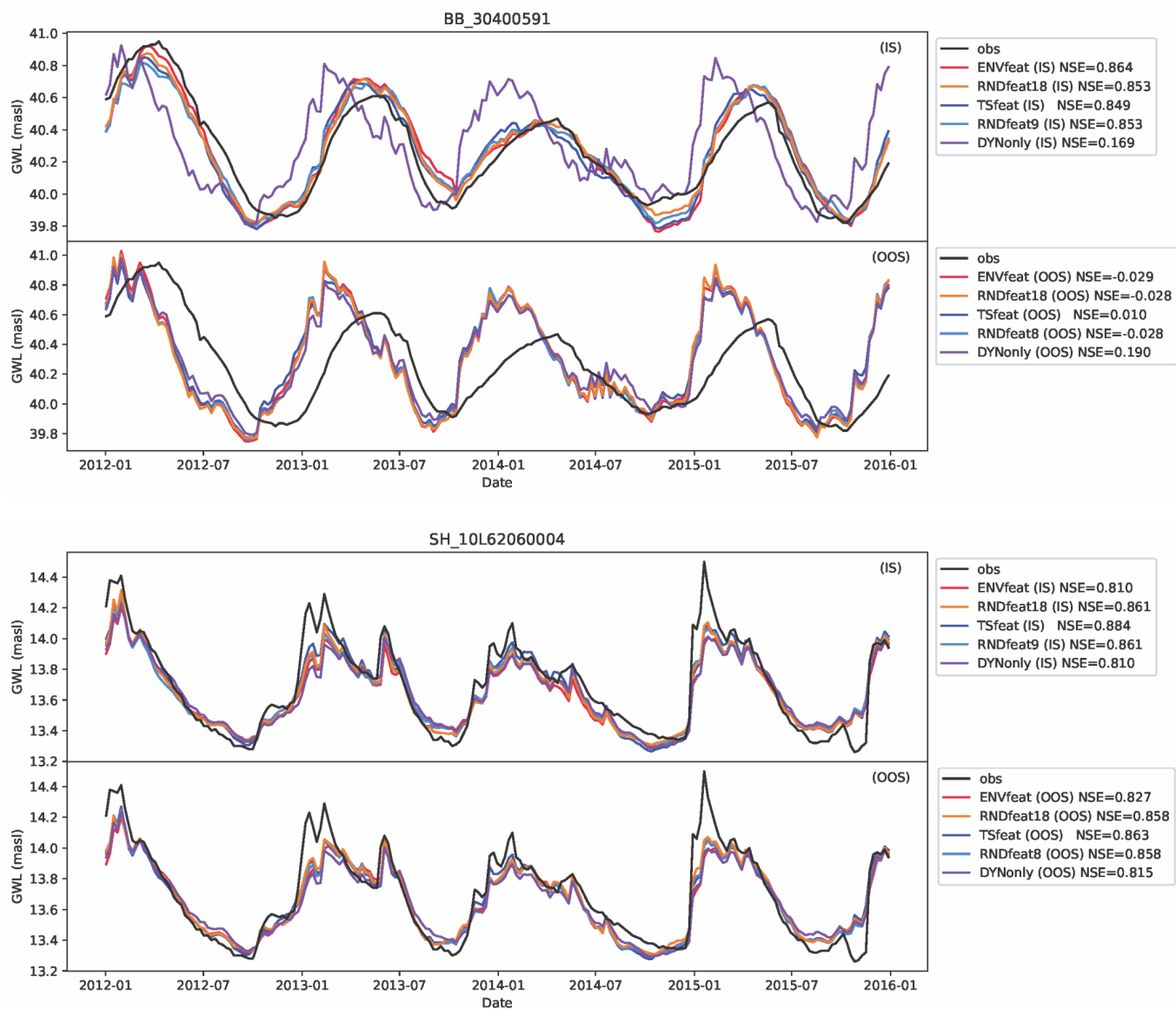
Figure 5. Examples of predictions of groundwater levels in the test period (2012-2015) for two wells, representing groups of similar behaviour (see text for details).

4.2 Permutation Feature Importance

By looking at the models' input feature importance, further insights can be gained. We applied Permutation Feature Importance to detect the relative importance of individual input features in the trained models. Figure 6 shows the feature importance of the ENVfeat, TSfeat and RNDfeat models, separately for static and dynamic features (dynamic feature importance includes all model variants).

The first thing to note is that every individual dynamic feature is much more important than any of the static features, as the permuted MSE increase is higher by orders of magnitude for dynamic features. Thereby, as expected, precipitation and temperature are the most important features, followed by the sinus curve fitted to temperature. Relative humidity is least important. Even though there is some instability involved in these results, with especially Tsin and rH, but also P and T experiencing heavy outliers up to 2 magnitudes above the median, by and large this confirms the findings of previous studies (Wunsch et al., 2021b). Admittedly, a more stable feature ranking could be obtained with alternative methods like SHAP (Lundberg and Lee, 2017), which was, however, not applied due to limited computational resources available to the authors.

Among static features, we find much more indifference between individual features (figure 6). Among static environmental features, CLC land cover comes out on top. This seems plausible because of its conceptual importance, and because it is the only feature representing land cover (although comprising 11 categories), unifying all information of land cover forcings, i.e. being informationally dense for the model. However, all environmental features show about the same feature importance as their random counterparts, confirming conclusions drawn in section 4.1 that they do not contribute any meaningful and thus generalizable information to the model.

Static time series features are a bit more sensitive, with high pulse duration (HPD) and annual periodicity (P52) outpacing all other time series features by some margin. But feature importance remains on a very low level for all other time series features, which are even surpassed by the average (relative) importance of the nine random features, the exception being high pulse duration (HPD). This allows the conclusion that this is the only static feature that might provide some meaningful/generalizable information to the model. It also could be the reason, that the TSfeat model at least slightly outperforms the RNDfeat9 model (compare fig. 3).

In general, however, the most important finding to take from this result is the fact that all static features are orders of magnitude less important than the dynamic features, which implies that the model draws the majority of information used for prediction from the shared dynamic features. This can be used as an explanation for the finding that ENVfeat, TSfeat and RNDfeat models perform almost identically (see fig 3), while DYNonly is able to outperform them out-of-sample.

5 Conclusions

The results of our work allow two main conclusions. First, in the in-sample setting, entity-aware global models work well and their performance can keep up with those of single-well models. All proposed model variants reached slightly better scores than the state-of-the-art single station model. However, contrary to our initial hypothesis, the TSfeat model does not outperform the ENVfeat model. Moreover, the RNDfeat models - having random integers instead of "meaningful" static features - performed
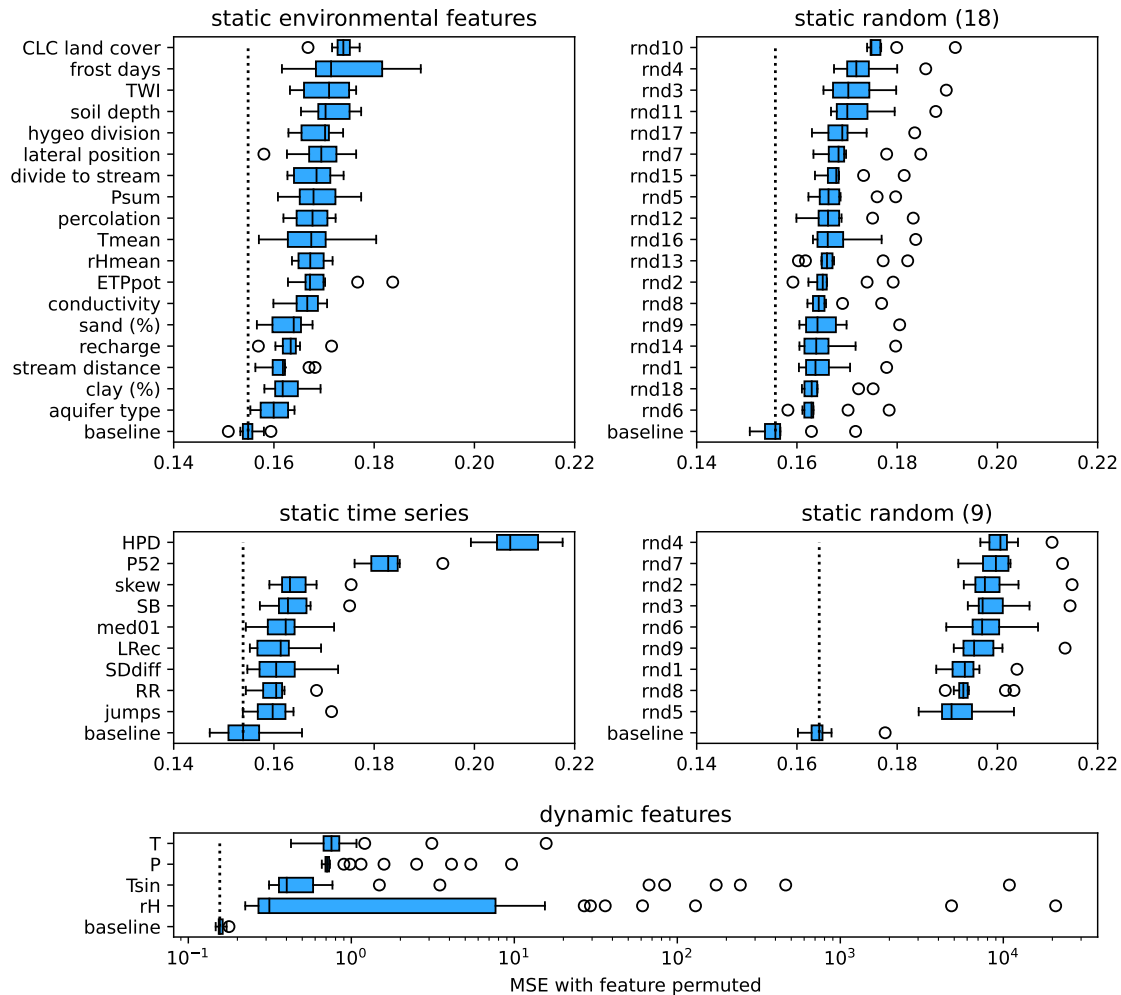
Figure 6. Permutation feature importance of the ENVfeat, TSfeat and RNDfeat models, separately for static and dynamic features (dynamic feature importance includes all model variants). The dotted lines indicate the MSE of the baseline model, which includes all features. Larger deviations from this baseline indicates higher feature importance.

equally well as the TSfeat and ENVfeat variants. Against the backdrop of the in-sample DYNonly model (trained without static features, only on meteorologic input), which had severely reduced performance, it is evident that this is because all tested sets of static features appear to only act as unique identifiers, enabling the model to differentiate between time series and memorizing their unique behaviour, but not to establish meaningful system-characterizing relationships based on the static features. Thus, we conclude that the models do not learn adequately from wells with similar information as provided by the static features. It may not be worth the effort to gather supposedly meaningful data, since random numbers might work just as well (as long as a decent number of random features is provided). This finding is in accordance with studies that have been

carried out in rainfall-runoff modelling (Li et al., 2022). Also, observed performance improvement over single-well models might just as well be due to architectural differences and the incorporation of additional dynamic input features (namely rH and Tsin) that weren't considered in the published single-well model results used as comparison. In other words, the models introduced here perform better, but not necessarily for the reason of being global or entity-aware, according to the commonly made claim that global models profit from additional similar data.

Second, out-of-sample performance of all model variants with static features expectedly decreases significantly. In general this is still a respectable performance, making a case for good generalizability in principle. However, the DYNonly model significantly outperformes TSfeat, ENVfeat and RNDfeat variants in this setting. This makes evident that static features, acting as unique identifiers in-sample, obscure learning of the only true meaningful or causal connection, namely of dynamic (meteorologic) input features to groundwater levels out-of-sample (i.e. when not included in training). In other words, the models are not able to learn that wells with a similar static feature combination should react similar to meteorological dynamic feature inputs in terms of groundwater level output. Instead, model skill is almost entirely based on learning from the dynamic input features. This might not come as a surprise for the environmental features, which were deemed to be afflicted with high uncertainties, but for the time-series features, since these proved in previous works to be well-suited to describe groundwater dynamics as a result of its reaction to meteorological inputs. Thus, our results suggest only a temporal generalizability potential - however valuable in itself - of entity-aware models, but lack evidence of true spatial generalizability potential - which remains the overarching aim of the field. However, this stands on the presented database, which, as stated above, might be too small, not diverse enough, and/or biased.

The tasks set by these conclusions are clear. First, since the dataset might just not contain enough data to take full advantage of global models, we plan to investigate this with a larger dataset, that covers groups of wells with several similarities as well as dissimilarities in a future study. The hypothesis is, that when more wells with similar meaningful static information are included in the dataset, the entity-aware model might then be able to better learn and generalize from the provided features. Second, our study revealed the glaring research gap of finding and compiling meaningful environmental descriptors of groundwater dynamics with true predictive power. The hydrogeologic discipline lacks large scale datasets of the kind. This severely restricts the development of hydrogeology as a machine learning research field, and the establishment of neural network models with physiographically meaningful internal structures, as was pursued in this study.

Code and data availability. The original groundwater level data are available free of charge from the respective local authorities: LUBW Baden-Wuerttemberg, LfU Bavaria, LfU Brandenburg, HLNUG Hesse, LUNG Mecklenburg-Western Pomerania, NLWKN Lower-Saxony, LANUV North Rhine-Westphalia, LfU Rhineland-Palatinate, LfULG Saxony, LHW Saxony-Anhalt and LLUR Schleswig-Holstein. With the kind permission of these local authorities, the processed groundwater level data have been published by Wunsch et al. (2021a). Meteorological input data was derived from the HYRAS dataset (Rauthe et al., 2013; Frick et al., 2014), which can be obtained free of charge for non-commercial purposes on request from the German Meteorological Service (DWD). The Python code as well as the underlying data is publicly accessible via GitHub: https://github.com/KITHydrogeology/2023-global-model-germany.
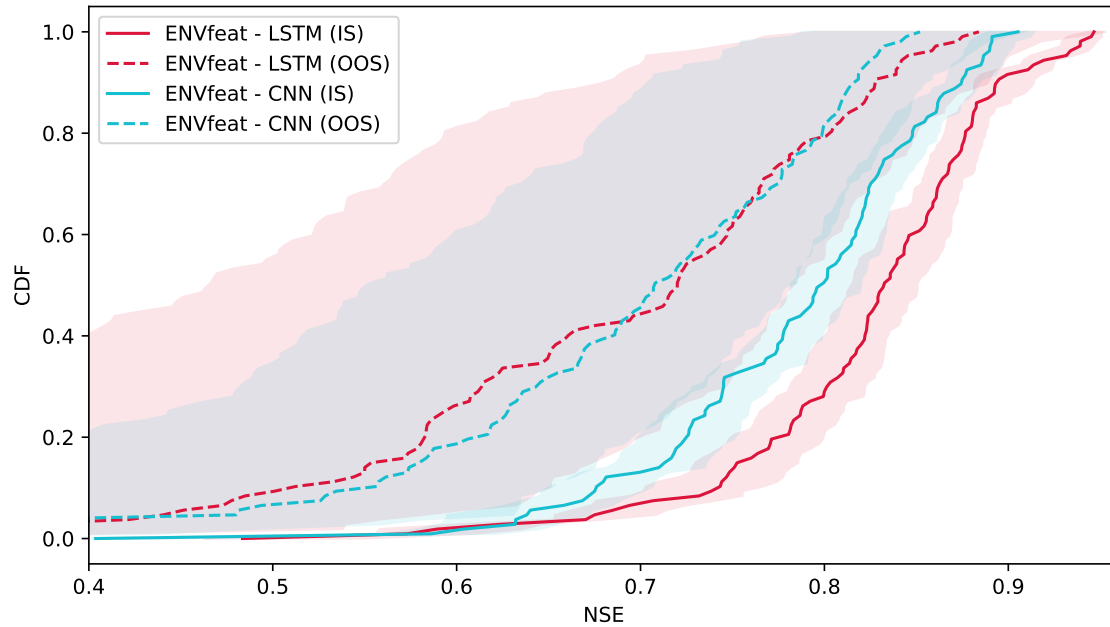
Figure A1. Comparing the performance of LSTM to CNN on the basis of the ENVfeat model variant. For the CNN model, the LSTM layer in the dynamic model thread is replaced with a CNN layer (followed by batchnorm and maxpool1D). The figure shows the models' performance in an in-sample (IS) and out-of-sample (OOS) setup. While CNNs and LSTMs perform almost equivalent in the OOS mode, CNNs are clearly inferior to LSTMs in the IS mode. Thus, LSTMs were used in this study.
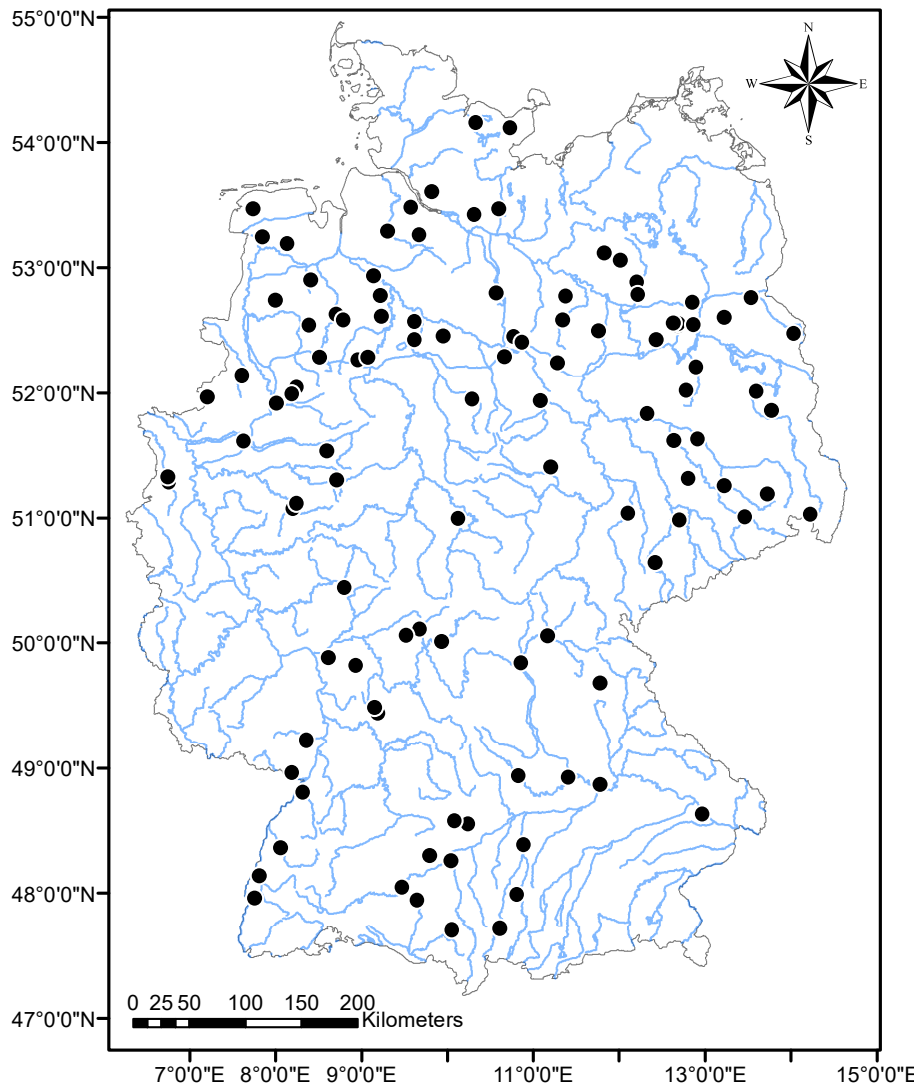
Figure A2. Location of the groundwater gauges within Germany.
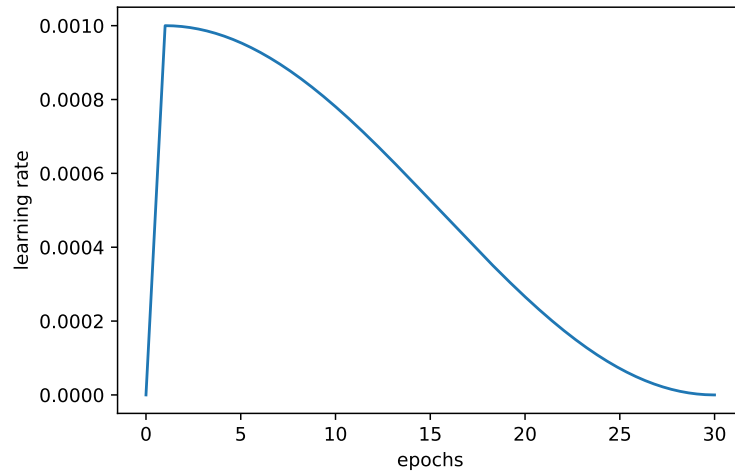
Figure A3. Visualisation of the learning rate schedule used in this study. It consists of one warmup epoch where the learning rate linearly increases from 0 to 0.001, followed by 29 epochs of cosine-shaped learning rate decline.
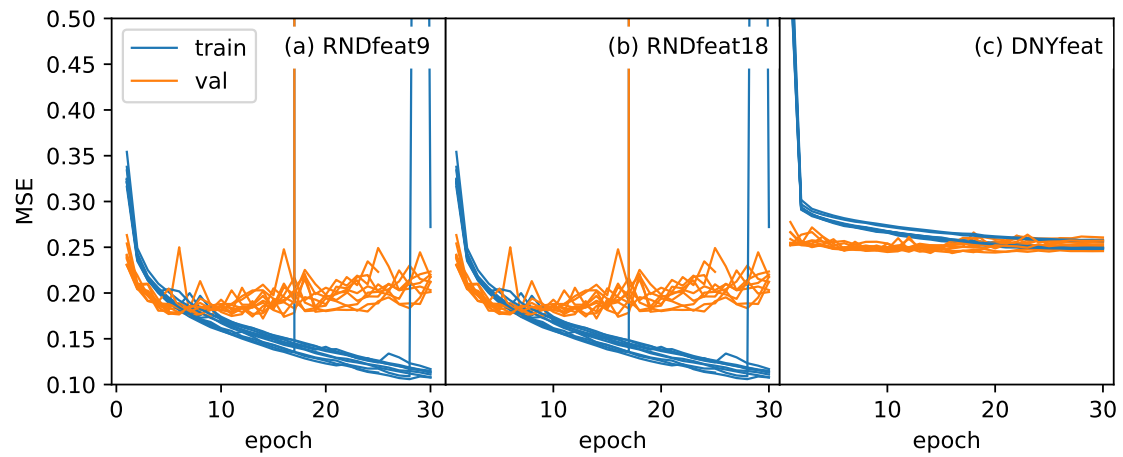
Figure A4. Losscurves of the training and validation period for the 10 different seed initialization runs for the (a) RNDfeat9 model, (b) RNDfeat18 model and (c) DYNonly model.
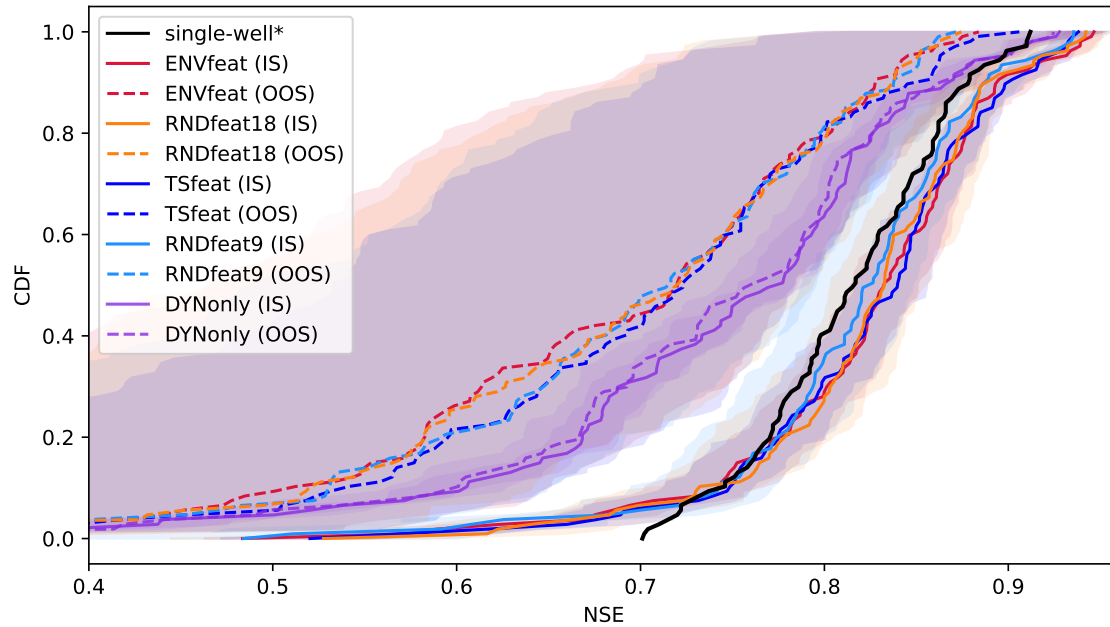
Figure A5. Cumulative distribution function of NSE of the model variants ENVfeat, TSfeat, RNDfeat and DYNonly in in-sample mode (IS) and out-of-sample mode (OOS) against the performance of the single-well models by Wunsch et al. (2022a) (*). Lines represent sorted median NSE scores of 10 ensemble members, envelopes represent ranges of the ensemble forecasts excluding the worst and best member.

Author contributions. BH and TL conceptualized the study, wrote the code, validated and visualized the results, and wrote the original paper draft. All three authors contributed to the methodology and performed review and editing tasks. TL supervised the work.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. All programming was done in Python version 3.9 (Van Rossum and Drake Jr, 1995) and the associated libraries, in-

465 cluding NumPy (Harris et al., 2020), Pandas (McKinney et al., 2010), Tensorflow (Abadi et al., 2016), Keras (Chollet et al., 2015), SciPy (Virtanen et al., 2020), Scikit-learn (Pedregosa et al., 2011) and Matplotlib (Hunter, 2007).

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: A system for large-scale machine learning, in: 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), pp. 265–283, 2016.

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, Hydrol. Earth Syst. Sci., 2017.

Ahamed, A., Knight, R., Alam, S., Pauloo, R., and Melton, F.: Assessing the utility of remote sensing data to accurately estimate changes in groundwater storage, Science of The Total Environment, 807, 150 635, 2022.

Barthel, R.: HESS Opinions "Integration of groundwater and surface water research: an interdisciplinary problem?", Hydrology and Earth System Sciences, 18, 2615–2628, https://doi.org/10.5194/hess-18-2615-2014, 2014.

Bedi, S., Samal, A., Ray, C., and Snow, D.: Comparative evaluation of machine learning models for groundwater quality assessment, Environmental Monitoring and Assessment, 192, 1–23, 2020.

Bengio, Y., Simard, P., and Frasconi, P.: Learning long-term dependencies with gradient descent is difficult, IEEE transactions on neural networks, 5, 157–166, 1994.

Beven, K. and Kirkby, M.: A physically based, variable contributing area model of basin hydrology, Hydrological Scienes Journal, 24, 43–69, 1979.

BGR: Soil map of Germany 1:5,000,000 (BUEK5000). Digital map data by the Federal Institute for Geosciences and Natural Resources (BGR), Hannover, 2005.

BGR: Soil map of Germany 1:200,000 (BUEK200). Digital map data by the Federal Institute for Geosciences and Natural Resources (BGR), Hannover, 2018.

BGR (Federal Institute for Geosciences and Natural Resources): Mean annual percolation rate from soil of Germany, 1:1,000,000 (SWR1000). Digital map data v1.0, Hannover., 2003.

BGR (Federal Institute for Geosciences and Natural Resources): Mean annual groundwater recharge of Germany 1961-1990, 1:1,000,000 (GWN1000). Digital map data v1.0, Hannover., 2019.

BGR (Federal Institute for Geosciences and Natural Resources) and SGD (German Federal States Geological Surveys): Hydrogeological spatial structure of Germany (HYRAUM). Digital map data v3.2, Hannover., 2015.

BGR (Federal Institute for Geosciences and Natural Resources) and SGD (German Federal States Geological Surveys): Hydrogeological map of Germany 1:250,000 (HUEK250). Digital map data v1.0.3. Hannover., 2019.

Bierkens, M. F. and Wada, Y.: Non-renewable groundwater use and groundwater depletion: a review, Environmental Research Letters, 14, 063 002, 2019.

Chollet, F. et al.: Keras, https://github.com/fchollet/keras, 2015.

Clark, S. R., Pagendam, D., and Ryan, L.: Forecasting Multiple Groundwater Time Series with Local and Global Deep Learning Networks, International Journal of Environmental Research and Public Health, 19, 5091, 2022.

Das, A., Kong, W., Leach, A., Sen, R., and Yu, R.: Long-term Forecasting with TiDE: Time-series Dense Encoder, arXiv preprint arXiv:2304.08424, 2023.

DWD Climate Data Center (CDC): Monatliche Raster der Summe der potentiellen Evapotranspiration über Gras, Version 0.x, 2023.

European Union, Copernicus Land Monitoring Service 2018, European Environment Agency (EEA): Corine Land Cover, 2018.

Famiglietti, J. S.: The global groundwater crisis, Nature Climate Change, 4, 945–948, 2014.

505 Fisher, A., Rudin, C., and Dominici, F.: All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously., J. Mach. Learn. Res., 20, 1–81, 2019.

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

510 Ghosh, R., Yang, H., Khandelwal, A., He, E., Renganathan, A., Sharma, S., Jia, X., and Kumar, V.: Entity Aware Modelling: A Survey, http://arxiv.org/abs/2302.08406, arXiv:2302.08406 [cs, stat], 2023.

Gleeson, T., Wagener, T., Döll, P., Zipper, S. C., West, C., Wada, Y., Taylor, R., Scanlon, B., Rosolem, R., Rahman, S., et al.: GMD Perspective: The quest to improve the evaluation of groundwater representation in continental to global scale models, Geoscientific Model Development Discussions, 2021, 1–59, 2021.

515 Goodfellow, I., Bengio, Y., and Courville, A.: Deep Learning, MIT Press, http://www.deeplearningbook.org, 2016.

Haaf, E., Giese, M., Heudorfer, B., Stahl, K., and Barthel, R.: Physiographic and Climatic Controls on Regional Groundwater Dynamics, Water Resources Research, 56, https://doi.org/10.1029/2019WR026545, 2020.

Haaf, E., Giese, M., Reimann, T., and Barthel, R.: Data-driven Estimation of Groundwater Level Time-Series at Unmonitored Sites Using Comparative Regional Analysis, Water Resources Research, p. e2022WR033470, 2023.

520 Haggerty, R., Sun, J., Yu, H., and Li, Y.: Application of machine learning in groundwater quality modeling-A comprehensive review, Water Research, p. 119745, 2023.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., Fernández del Río, J., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E.: Array programming with NumPy,

525 Nature, 585, 357–362, https://doi.org/10.1038/s41586-020-2649-2, 2020.

Hengl, T., Mendes De Jesus, J., Heuvelink, G. B. M., Ruiperez Gonzalez, M., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., and Kempen, B.: SoilGrids250m: Global gridded soil information based on machine learning, PLOS ONE, 12, e0169 748, https://doi.org/10.1371/journal.pone.0169748, 2017.

530 Heudorfer, B., Haaf, E., Stahl, K., and Barther, R.: Index-based characterization and quantification of groundwater dynamics, Water Resources Research, 55, 5575–5592, https://doi.org/10.1029/2018WR024418, 2019.

Hochreiter, S. and Schmidhuber, J.: Long short-term memory, Neural computation, 9, 1735–1780, 1997.

Hoelting, B. and Coldewey, W.: Hydrogeologie. Einführung in die Allgemeine und Angewandte Hydrogeologie., Springer Spektrum, 8 edn., 2013.

535 Hrachowitz, M., Savenije, H., Bloeschl, G., McDonnell, J., Sivapalan, M., Pomeroy, J., Arheimer, B., Blume, T., Clark, M., and Ehret, U.: A decade of Predictions in Ungauged Basins (PUB)—a review, Hydrological sciences journal, 58, 1198–1255, 2013.

Hunter, J. D.: Matplotlib: A 2D graphics environment, Computing in science & engineering, 9, 90–95, 2007.

Kingma, D. P. and Ba, J.: Adam: A Method for Stochastic Optimization, http://arxiv.org/abs/1412.6980, arXiv:1412.6980 [cs], 2014.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M.: Rainfall–runoff modelling using Long Short-Term Memory (LSTM)

540 networks, Hydrology and Earth System Sciences, 22, 6005–6022, https://doi.org/10.5194/hess-22-6005-2018, 2018.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., and Nearing, G. S.: Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning, Water Resources Research, 55, 11 344–11 354, https://doi.org/10.1029/2019WR026065, 2019a.

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, Hydrology and Earth System Sciences, 23, 5089–5110, https://doi.org/10.5194/hess-23-5089-2019, 2019b.

Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., Nevo, S., Shalev, G., and Matias, Y.: Caravan - A global community dataset for large-sample hydrology, Scientific Data, 10, 61, https://doi.org/10.1038/s41597-023-01975-w, 2023.

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 Resources Research, 58, https://doi.org/10.1029/2021WR031794, 2022.

Li, Y., Wei, C., and Ma, T.: Towards Explaining the Regularization Effect of Initial Large Learning Rate in Training Neural Networks, Advances in Neural Information Processing Systems, 32, 2019.

Linke, S., Lehner, B., Ouellet Dallaire, C., Ariwi, J., Grill, G., Anand, M., Beames, P., Burchard-Levine, V., Maxwell, S., Moidu, H., Tan, F., and Thieme, M.: Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution, Scientific Data, 6, 283, https://doi.org/10.1038/s41597-019-0300-6, 2019.

Liu, Q., Yang, M., Mohammadi, K., Song, D., Bi, J., and Wang, G.: Machine Learning Crop Yield Models Based on Meteorological Features and Comparison with a Process-Based Model, Artificial Intelligence for the Earth Systems, 1, e220 002, 2022.

Lundberg, S. M. and Lee, S.-I.: A unified approach to interpreting model predictions, Advances in neural information processing systems, 30, 2017.

Majumdar, S., Smith, R., Butler Jr, J., and Lakshmi, V.: Groundwater withdrawal prediction using integrated multitemporal remote sensing data sets and machine learning, Water Resources Research, 56, e2020WR028 059, 2020.

Majumdar, S., Smith, R., Conway, B. D., and Lakshmi, V.: Advancing remote sensing and machine learning-driven frameworks for groundwater withdrawal estimation in Arizona: Linking land subsidence to groundwater withdrawals, Hydrological Processes, 36, e14 757, 2022.

McKinney, W. et al.: Data structures for statistical computing in python, in: Proceedings of the 9th Python in Science Conference, vol. 445, pp. 51–56, Austin, TX, 2010.

Miebs, G., Mochol-Grzelak, M., Karaszewski, A., and Bachorz, R. A.: Efficient strategies of static features incorporation into the recurrent neural network, Neural Processing Letters, 51, 2301–2316, 2020.

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in Python, Journal of machine learning research, 12, 2825–2830, 2011.

Rajaee, T., Ebrahimi, H., and Nourani, V.: A review of the artificial intelligence methods in groundwater level modeling, Journal of hydrology, 572, 336–351, 2019.

Ransom, K. M., Nolan, B. T., Stackelberg, P., Belitz, K., and Fram, M. S.: Machine learning predictions of nitrate in groundwater used for drinking supply in the conterminous United States, Science of the Total Environment, 807, 151 065, 2022.

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

Richter, B., Baumgartner, J., Powell, J., and Braun, D.: A method for assessing hydrologic alteration within ecosystems, Conservation Biology, 10, 1163–1174, 1996.

Sivapalan, M., Takeuchi, K., Franks, S., Gupta, V., Karambiri, H., Lakshmi, V., Liang, X., McDonnell, J., Mendiondo, E., and O'Connell, P.: IAHS Decade on Predictions in Ungauged Basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences, Hydrological Sciences Journal, 48, 857–880, 2003.

Smith, L. N. and Topin, N.: Super-Convergence: Very Fast Training of Neural Networks Using Large Learning Rates, http://arxiv.org/abs/1708.07120, arXiv:1708.07120 [cs, stat], 2018.

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.

Van Rossum, G. and Drake Jr, F. L.: Python reference manual, Centrum voor Wiskunde en Informatica Amsterdam, 1995.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors: SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, Nature Methods, 17, 261–272, https://doi.org/10.1038/s41592-019-0686-2, 2020.

Wada, Y., Van Beek, L. P., Van Kempen, C. M., Reckman, J. W., Vasak, S., and Bierkens, M. F.: Global depletion of groundwater resources, Geophysical research letters, 37, 2010.

Wunsch, A. and Liesch, T.: Entwicklung und Anwendung von Algorithmen zur Berechnung von Grundwasserständen an Referenzmessstellen auf Basis der Methode Künstlicher Neuronaler Netze, 2020.

Wunsch, A., Liesch, T., and Broda, S.: Weekly groundwater level time series dataset for 118 wells in Germany, https://doi.org/10.5281/zenodo.4683879, 2021a.

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), Hydrology and Earth System Sciences, 25, 1671–1687, https://doi.org/10.5194/hess-25-1671-2021, 2021b.

Wunsch, A., Liesch, T., and Broda, S.: Deep learning shows declining groundwater levels in Germany until 2100 due to climate change, Nature Communications, 13, 1221, 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 Resources Management, 36, 39–54, https://doi.org/10.1007/s11269-021-03006-y, 2022b.

WWAP: The United Nations world water development report 2015: water for a sustainable world, UN World Water Assessment Programme, 2015.

Zeng, A., Chen, M., Zhang, L., and Xu, Q.: Are transformers effective for time series forecasting?, in: Proceedings of the AAAI conference on artificial intelligence, vol. 37, pp. 11 121–11 128, 2023.