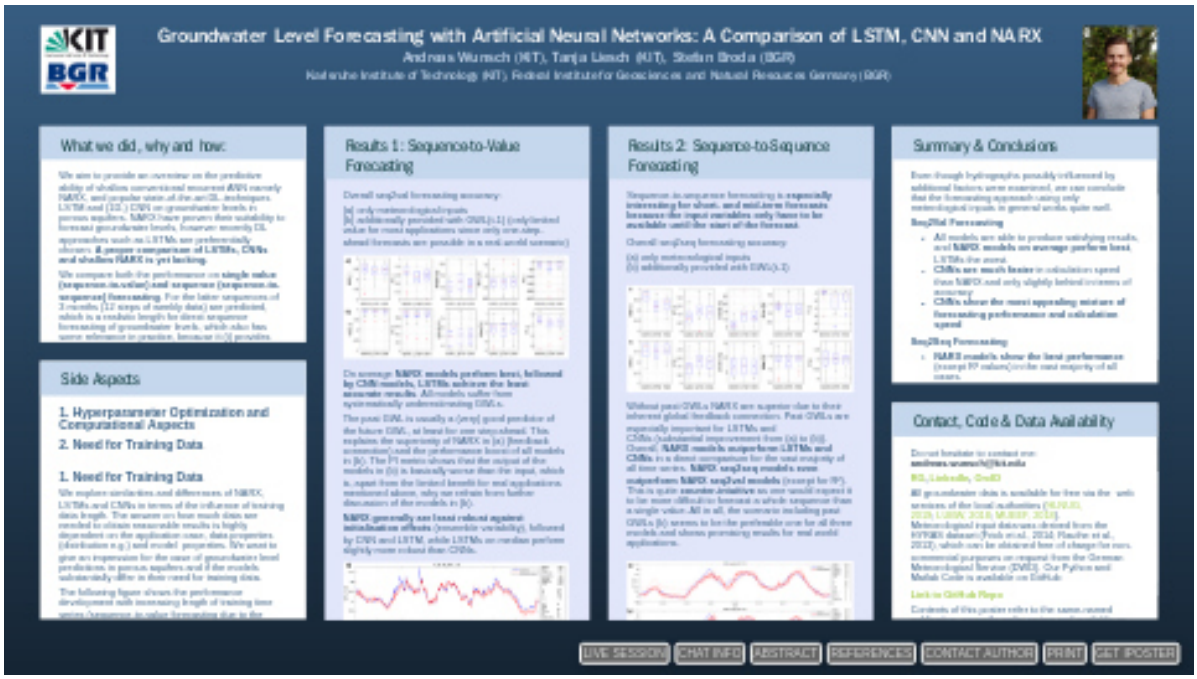
Groundwater Level Forecasting with Artificial Neural Networks: A Comparison of LSTM, CNN and NARX



Andreas Wunsch (KIT), Tanja Liesch (KIT), Stefan Broda (BGR)

Karlsruhe Institute of Technology (KIT), Federal Institute for Geosciences and Natural Resources Germany (BGR)

PRESENTED AT:

WHAT WE DID, WHY AND HOW:

We aim to provide an overview on the predictive ability of shallow conventional recurrent ANN namely NARX, and popular state-of-the-art DL-techniques LSTM and (1D-) CNN on groundwater levels in porous aquifers. NARX have proven their suitability to forecast groundwater levels, however recently DL approaches such as LSTMs are preferentially chosen. **A proper comparison of LSTMs, CNNs and shallow NARX is yet lacking.**

We compare both the performance on **single value (sequence-to-value) and sequence (sequence-to-sequence) forecasting**. For the latter sequences of 3 months (12 steps of weekly data) are predicted, which is a realistic length for direct sequence forecasting of groundwater levels, which also has some relevance in practice, because it (i) provides useful information for many decision-making applications (e.g. groundwater management), and (ii) is also an established time-span in meteorological forecasting, known as seasonal forecasts. This ensures applicability in a real world scenario. We use data from 17 groundwater wells within the Upper Rhine Graben region in Germany and France, selected based on prior knowledge and representing the full bandwidth of groundwater dynamic types in the region. Further we use only **widely available and easy to measure meteorological input parameters** (precipitation, temperature and relative humidity), which makes our approach widely applicable.

We further explore computational aspects of all models during hyperparmeter optimization as well as compare their need for triaining data.

Models

We use the following models:

- **Nonlinear Autoregressive Exogenous Model (NARX)**
- **Long Short-Term Memory (LSTM)**
- **Convolutional Neural Networks (CNN)**

Nonlinear autoregressive networks with exogenous input are a specific type of recurrent neural networks (RNNs) that extend the well-known structure of feed-forward multilayer perceptrons (MLP) by a global feedback connection between output and input layer. NARX also contain a short-term memory, i.e. delay vectors for each input (and feedback), which allow the availability of several input time steps simultaneously, depending on the length of the vector.

Long Short-Term Memory networks are recurrent neural networks which are widely applied to model sequential data like time series or natural language. LSTMs can remember long-term dependencies because they have been explicitly designed to overcome the problem of vanishing gradients. Besides the hidden state of RNNs, LSTMs have a cell memory (or cell state) to store information and three gates (forget, input, output) to control information flow.

CNNs are neural networks, which are predominantly used for image recognition and classification. However, they work also well on signal processing tasks (NLP for example). CNNs usually comprise three different layers: convolutional layers (filters and feature maps), pooling layers (down-sampling) and fully connected dense layers.

Input Parameters

We use:

- **precipitation (P)**
- **temperature (T)**
- **relative humidity (rH)**

These are widely available and easy to measure, which makes this approach easily transferable and thus applicable almost everywhere.

To provide the model with noise-free information on seasonality, which often allows significantly improved predictions to be made, we use:

- **a sinusoidal signal fitted to the temperature curve (Tsin)**,

as an additional synthetic input parameter.

A very good predictor is also the past GWL until one step before the start of the forecast (t):

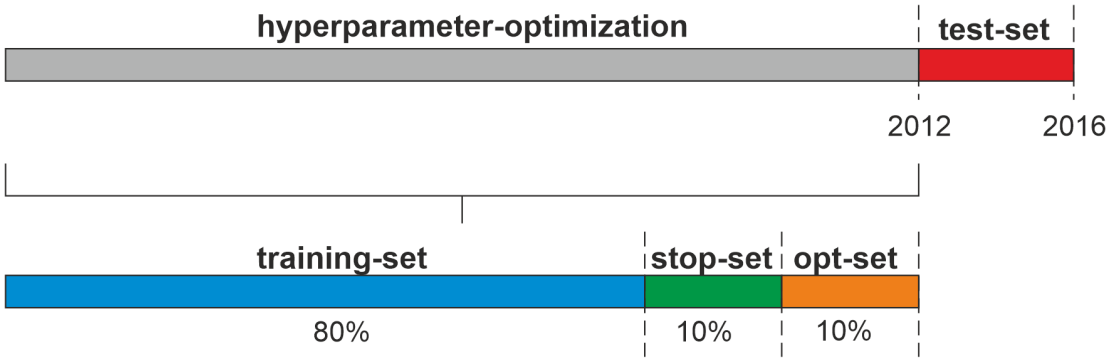
- **GWL(t-1)**

Depending on the purpose and methodological setup it does not always make sense to include this parameter; however, where meaningful we explored also past GWLs as Inputs.

Model calibration and evaluation

Hyperparameter Optimization is conducted by applying **Bayesian optimization** using the python implementation by Nogueira (2014) (https://github.com/fmfn/BayesianOptimization) and the built in Matlab optimization respectively. We apply 50 optimization steps as a minimum, using expected improvement as acquisition function. Strictly speaking, input selection is no hyperparameter optimization problem, however, **the algorithm can also be applied to select an appropriate set of inputs**. This assumption applies in our study also to LSTM and CNN models. Precipitation as the presumably most important input ist fixed and not optimised.

The testing or evaluation period in this study for all models are 4 years (2012 to 2016). The rest of the data is split into 80% for training, 10% for early stopping and 10% for testing during HP-Optimization (opt-set):



To minimize initialization influence we always calculate small ensembles with 5 different pseudo random seeds. For the final model evaluation in the test period (2012-2016) we use **10 pseudo-random initializations** and calculate errors of the median forecast.

We further calculate several metrics to judge accuracy: **Nash-Sutcliffe Efficiency (NSE), coefficient of determination (R^2), absolute and relative root mean squared error (RMSE/rRMSE), absolute and relative Bias (Bias/rBias) as well as Persistency index (PI)**.

The persistency index PI basically compares the performance to a naïve model that uses the last known observed groundwater level at the time the prediction starts (t). This is particularly important to judge the performance, when past groundwater levels (GWL(t-1)) are used as inputs, because especially in this case the model should outperform a naïve forecast (PI > 0).

Data & Study Area

We examine the groundwater level forecasting performance at **17 groundwater wells within the Upper Rhine Graben (URG) area**, the largest groundwater resource in central Europe. The wells are selected from a larger dataset from the region with more than 1800 hydrographs, based on prior knowledge to represent the full bandwidth of groundwater dynamics occurring in the dataset. The whole dataset mainly consists of shallow wells from the uppermost aquifer within the Quaternary sand/gravel sediments of the URG. The shortest time series starts in 1994, the longest in 1967, however, most hydrographs (12) start between 1980 and 1983. We exclusively consider **weekly time steps** for both groundwater and meteorological data.

SIDE ASPECTS

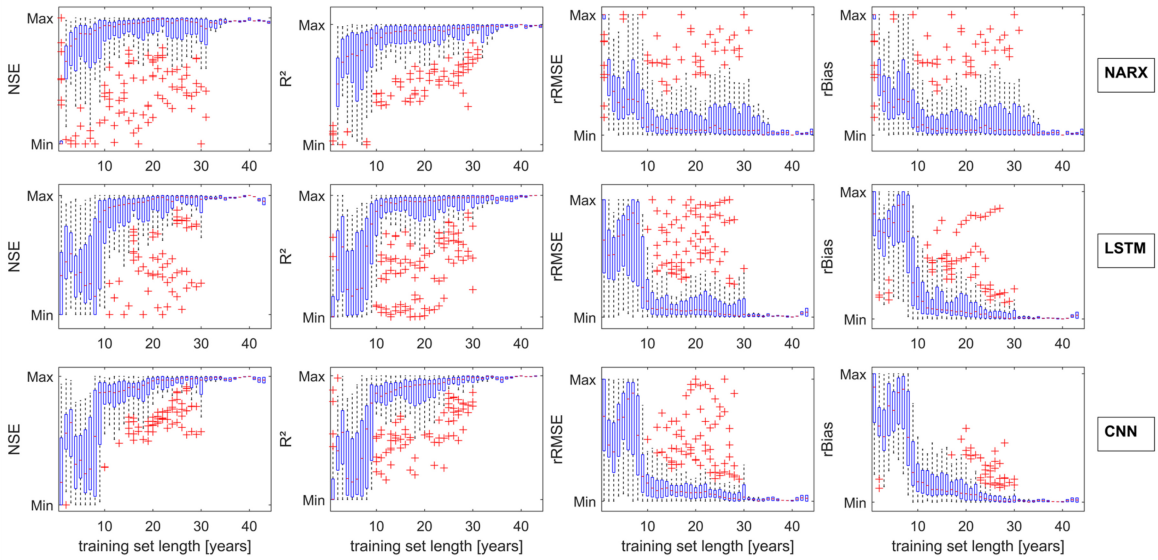
1. Hyperparameter Optimization and Computational Aspects

2. Need for Training Data

1. Need for Training Data

We explore similarities and differences of NARX, LSTMs and CNNs in terms of the influence of training data length. The answer on how much data are needed to obtain reasonable results is highly dependent on the application case, data properties (distribution e.g.) and model properties. We want to give an impression for the case of groundwater level predictions in porous aquifers and if the models substantially differ in their need for training data.

The following figure shows the performance development with increasing length of training time series (sequence-to-value forecasting due to the easier interpretability):



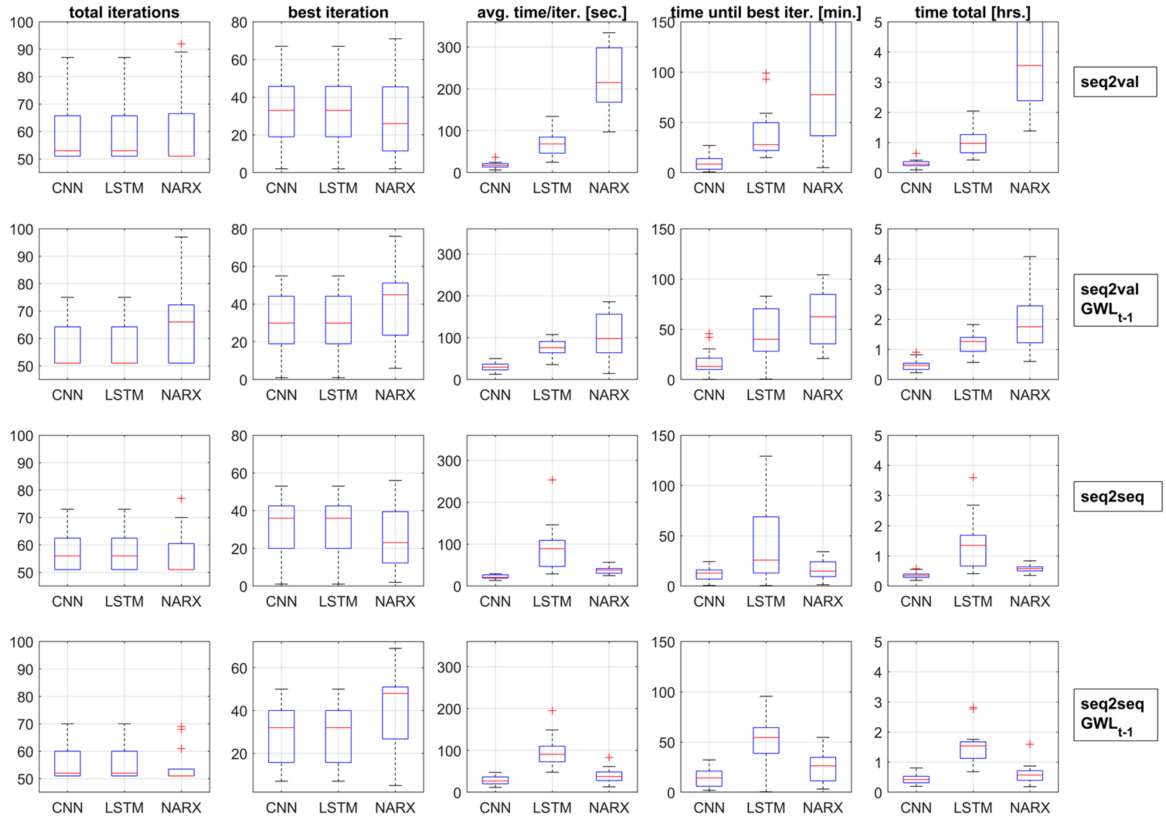
As expected, we observe significant improvements with additional training data. NARX models seem to improve somehow continuously, whereas **LSTMs and CNNs show some kind of threshold** (about 10 years) for a strongly improving performance. We explored the reason for this threshold and observed that when stopping the training systematically five years earlier, the threshold now correspondingly shifts. Additionally, several standard statistic values (mean, median, variance, several quantiles a.o.) show similar thresholds. Thus, the early years of the 2000s, seem to be especially relevant for our test period. **This is a highly dataset-specific observation that cannot be generalized;** however, this also shows that it is vital to include relevant training data, which is, however, not very easy to identify.

Nevertheless, as a rule of thumb the chance of using the right data, increases with the amount of available data. These findings are supported by the observation that not every additional year improves the accuracy, only the overall trend is positive. This seems plausible, because especially when conditions change over time, the models also can learn behaviour that is no longer valid and which possibly decreases future forecast performance. One should therefore not only include as much data as possible, but also carefully evaluate and also possibly shorten the training data base if necessary.

2. Hyperparameter Optimization and Computational Aspects

We used a standard desktop PC to build and train our models to give an realisitc impression on the computational performance of the different models and share practice relevant insights for fellow hydrogeologists. We trained CNNs and NARX on the CPU (AMD-Ryzen 9 3900X) and LSTMs on the GPU (Nvidia GeForce RTX 2070 Super). We chose the fastest option each. NARX were implemented in Matlab, both LSTMs and CNNs were implemented using Python 3.8.

Depending on the forecasting approach (seq2val/seq2seq) and available inputs (with/without past GWL), there were noticeable differences with regard to the number of iterations required for the hyperparameter optimization and the associated time needed:



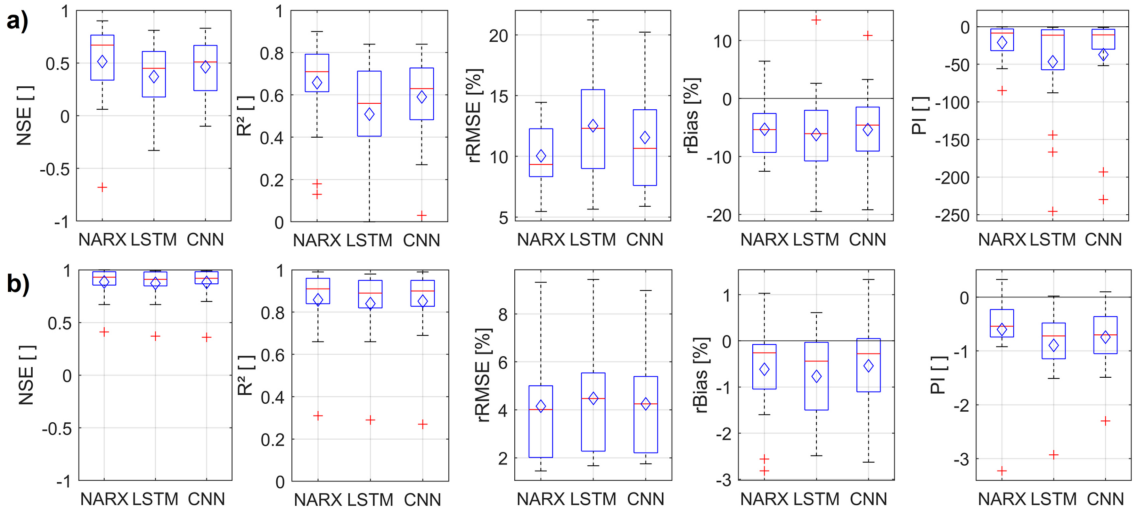
In the majority of cases the best iteration was found in less than 33 steps (col. 2), the minimum as well as the maximum number of iteration steps were therefore obviously sufficient. It is interesting that for CNN and LSTM the number of steps is similar throughout the experiments, whereas for NARX the inclusion of past GWLs as input caused an increase of iterations.

Columns three to five show substantial differences concerning the calculation speed of the three model types. **CNNs outperform all other models systematically**, however, concerning the sequence-2-sequence forecasts, NARX models speed up substantially compared to seq2val performance and can almost keep up with CNNs. We also observe that LSTMs seem to slow down when including GWL(t-1) as input or when performing seq2seq forecasts.

RESULTS 1: SEQUENCE-TO-VALUE FORECASTING

Overall seq2val forecasting accuracy:

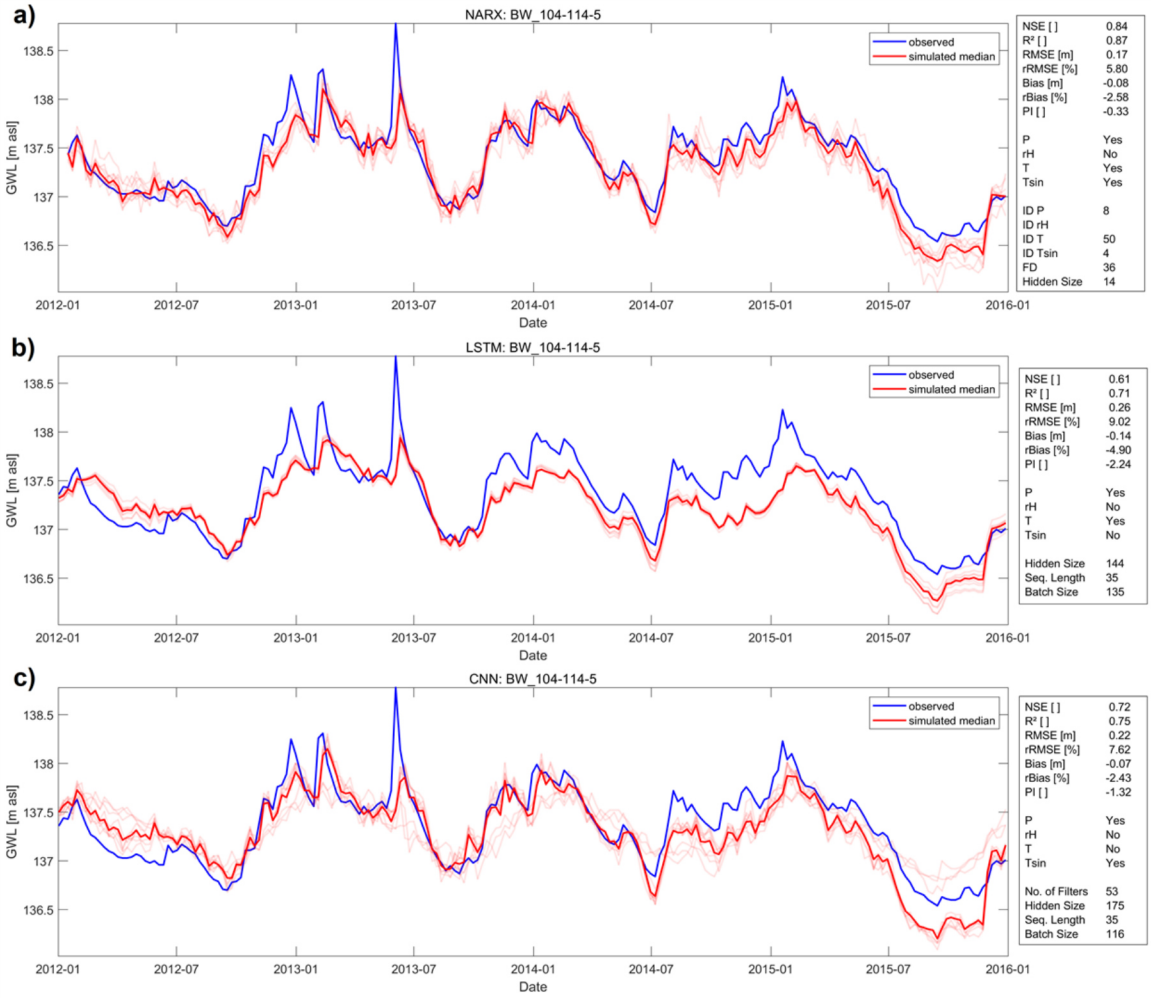
(a) only meteorological inputs
(b) additionally provided with GWL(t-1) (only limited value for most applications since only one-step-ahead forecasts are possible in a real-world scenario)



On average **NARX models perform best, followed by CNN models, LSTMs achieve the least accurate results**. All models suffer from systematically underestimating GWLs.

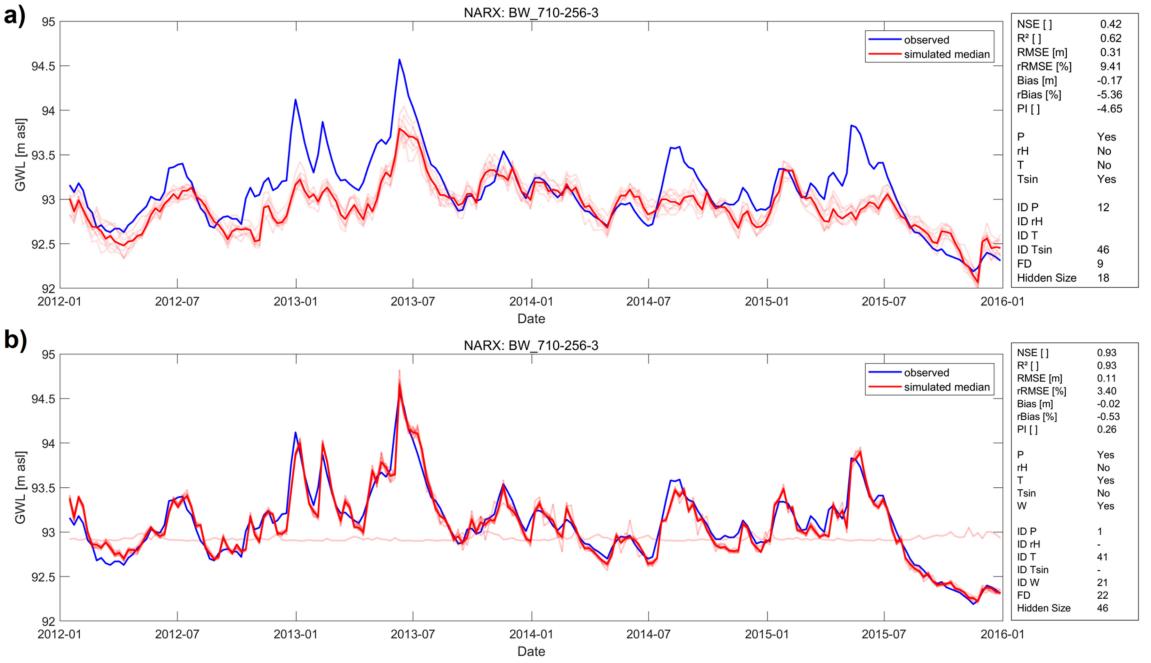
The past GWL is usually a (very) good predictor of the future GWL, at least for one step ahead. This explains the superiority of NARX in (a) (feedback connection) and the performance boost of all models in (b). The PI metric shows that the output of the models in (b) is basically worse than the input, which is, apart from the limited benefit for real applications mentioned above, why we refrain from further discussion of the models in (b).

NARX generally are least robust against initialisation effects (ensemble variability), followed by CNN and LSTM, while LSTMs on median perform slightly more robust than CNNs.



The figure above shows exemplarily the forecasting performance of all three models for well BW_104-114-5, where all models consistently achieved good results in terms of accuracy. The NARX model (a) outperforms both LSTM (b) and CNN (c) models and shows very high NSE and R^2 values between 0.8 and 0.9.

If groundwater dynamic is **significantly influenced by other factors than meteorology**, the figure below shows that including other relevant inputs can improve accuracy significantly (b). In this case Rhine River water levels (W) (major streamflow) were provided additionally. This also causes lower dependency to the model initialization, which corresponds also to other time-series, where we often find smaller influence the more relevant the input data is. Little accuracy of our approach is therefore probably often due to insufficient input data on a case-by-case basis, not necessarily because of an inadequate modelling approach.

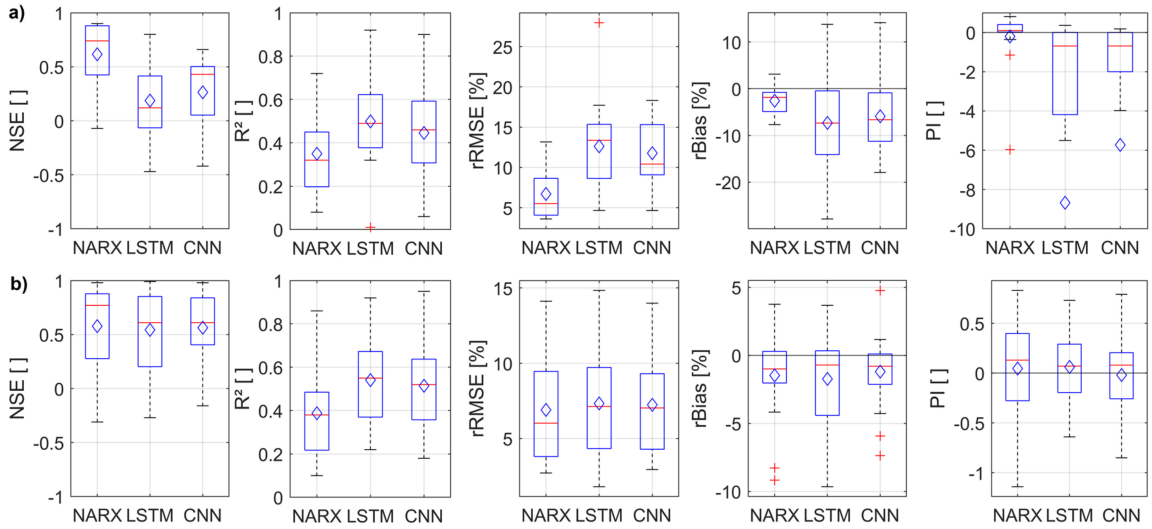


RESULTS 2: SEQUENCE-TO-SEQUENCE FORECASTING

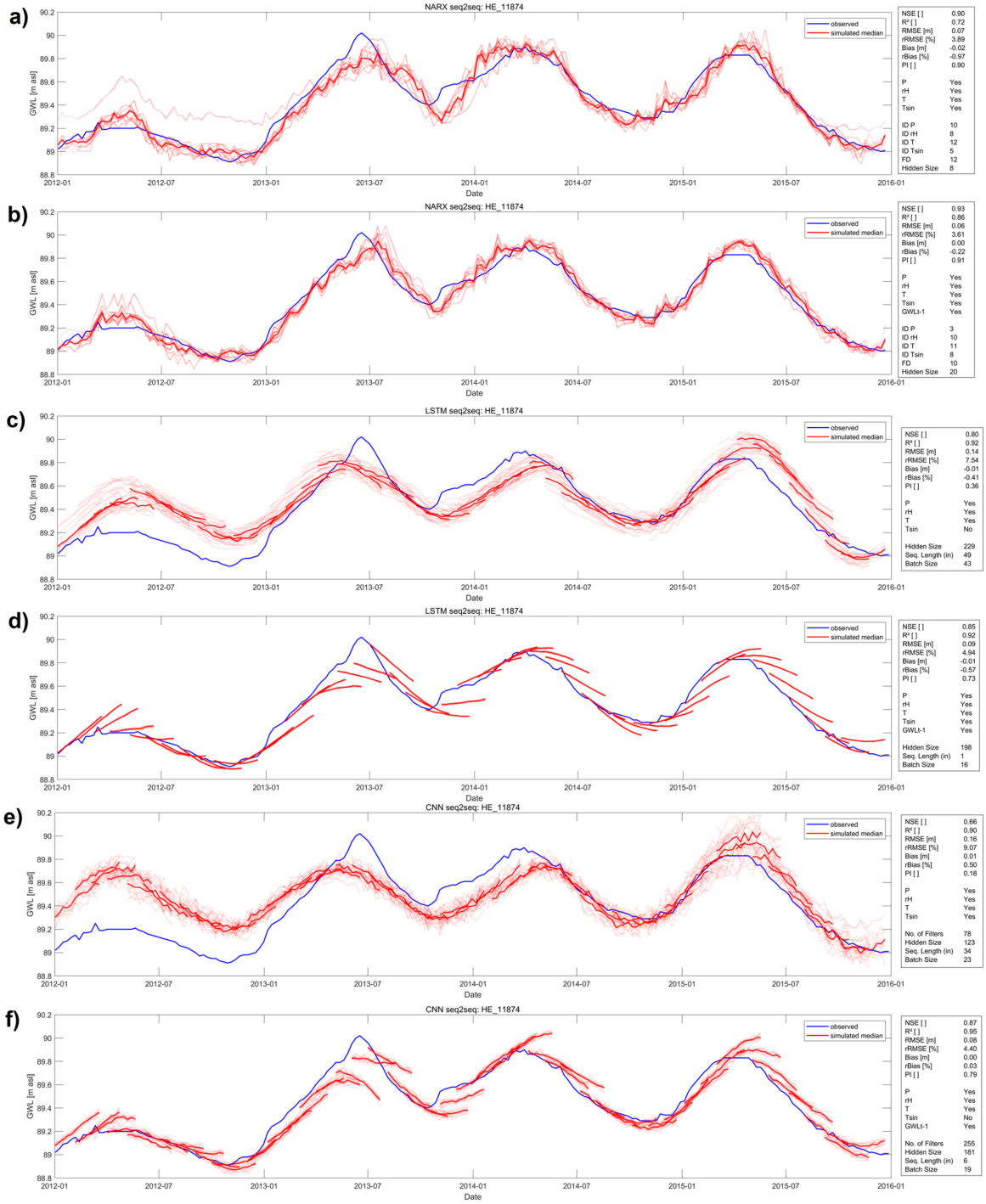
Sequence-to-sequence forecasting is **especially interesting for short- and mid-term forecasts because the input variables only have to be available until the start of the forecast**.

Overall seq2seq forecasting accuracy:

(a) only meteorological inputs
(b) additionally provided with GWL(t-1)



Without past GWLs NARX are superior due to their inherent global feedback connection. Past GWLs are especially important for LSTMs and CNNs (substantial improvement from (a) to (b)). Overall, **NARX models outperform LSTMs and CNNs** in a direct comparison for the vast majority of all time series. **NARX seq2seq models even outperform NARX seq2val models** (except for R^2). This is quite **counter-intuitive** as one would expect it to be more difficult to forecast a whole sequence than a single value. All in all, the scenario including past GWLs (b) seems to be the preferable one for all three models and shows promising results for real world applications.



The figures above summarise exemplarily for well HE_11874 the sequence-to-sequence forecasting performance for:

- NARX (a,b), LSTMs (c,d), CNNs (e,f),
- only meteorological inputs (a,c,e)
- also with past GWL inputs (b,d,f).

As stated above GWL(t-1)-input substantially improves the performance of LSTMs and CNNs, however, NARX forecasts in this case only improve very slightly. Especially for LSTMs and CNNs it is easily visible that the sequence forecasts of the better models (d,f) mostly estimate the intensity of a future groundwater level change too conservatively (extreme values are typically under-represented in the distribution of the training data).

The initialization dependency of LSTMs and CNNs is significantly lower than for NARX, with LSTMs being even more robust than CNNs. Despite the **significantly lower robustness of NARX models** the median ensemble **nevertheless is of high accuracy.** All models, but especially NARX models, therefore should not be evaluated without including an initialization ensemble.

Sequence predictions of NARX models overlap exactly in contrary to CNN and LSTM forecasts. The reason for this is the **differing technical approach for seq2seq forecasting**. While LSTMs and CNNs use multiple output neurons to predict multiple steps at once, this approach for us did not yield meaningful results in case of NARX, probably because of feedback connection issues. Instead we used one NARX output neuron to predict a multi-element vector at once.

SUMMARY & CONCLUSIONS

Even though hydrographs possibly influenced by additional factors were examined, we can conclude that the forecasting approach using only meterological inputs in general works quite well.

Seq2Val Forecasting

- All models are able to produce satisfying results, and **NARX models on average perform best**, LSTMs the worst.
- **CNNs are much faster** in calculation speed than NARX and only slightly behind in terms of accuracy
- **CNNs show the most appealing mixture of forecasting performance and calculation speed**

Seq2Seq Forecasting

- **NARX models show the best performance** (except R^2 values) in the vast majority of all cases.
- A **strong speed up of NARX in calculation time** compared to Seq2Val experiments makes NARX the preferable model for Seq2Seq predictions
- CNNs and LSTMs are significantly more least robust against initialisation effects, which nevertheless can be easily handled (also for NARX) by implementing a forecasting ensemble.

Need for Training Data

- As expected, we found that in principle the longer the training data, the better.
- A noteworthy threshold seems to exist for about 10 years of weekly training data, below which the performance becomes significantly worse (especially for CNN and LSTM), however, we fount this threshold to be highly dataset specific.

Typical groundwater level forecasting scenarios do not contribute as much data as would probably be needed for the DL approaches to significantly outperform the shallow NN. The latter should therefore not be neglected in model approach selection processes, due to the more recent and thus probably more appealing DL approaches.

CONTACT, CODE & DATA AVAILABILITY

Do not hesitate to contact me: **andreas.wunsch@kit.edu**

RG (https://www.researchgate.net/profile/Andreas_Wunsch4), LinkedIn (https://www.linkedin.com/in/andreaswunsch/), OrcID (https://orcid.org/0000-0002-0585-9549)

All groundwater data is available for free via the web services of the local authorities (HLNUG, 2019 (http://gruschu.hessen.de); LUBW, 2018 (http://udo.lubw.baden-wuerttemberg.de/public/); MUEEF, 2018 (https://geoportal-wasser.rlp-umwelt.de/)). Meteorological input data was derived from the HYRAS dataset (Frick et al., 2014; Rauthe et al., 2013), which can be obtained free of charge for non-commercial purposes on request from the German Meteorological Service (DWD). Our Python and Matlab Code is available on GitHub:

Link to GitHub Repo (https://github.com/AndreasWunsch/Groundwater-Level-Forecasting-with-ANNs-A-Comparison-of-LSTM-CNN-and-NARX)

Contents of this poster refer to the same-named publication, currently under review and available as preprint here: **Link to Preprint** (https://hess.copernicus.org/preprints/hess-2020-552/)

ABSTRACT

It is now well established to use shallow artificial neural networks (ANN) to obtain accurate and reliable groundwater level forecasts, which are an important tool for sustainable groundwater management. However, we observe an increasing shift from conventional shallow ANNs to state-of-the-art deep learning (DL) techniques, but a direct comparison of the performance is often lacking. Although they have already clearly proven their suitability, especially shallow recurrent networks frequently seem to be excluded from the study design despite the euphoria about new DL techniques and its successes in various disciplines. Therefore, we aim to provide an overview on the predictive ability in terms of groundwater levels of shallow conventional recurrent ANN namely nonlinear autoregressive networks with exogenous inputs (NARX), and popular state-of-the-art DL-techniques such as long short-term memory (LSTM) and convolutional neural networks (CNN). We compare both the performance on sequence-to-value (seq2val) and sequence-to-sequence (seq2seq) forecasting on a 4-year period, while using only few, widely available and easy to measure meteorological input parameters, which makes our approach widely applicable. We observe that for seq2val forecasts NARX models on average perform best, however, CNNs are much faster and only slightly worse in terms of accuracy. For seq2seq forecasts, mostly NARX outperform both DL-models and even almost reach the speed of CNNs. However, NARX are the least robust against initialization effects, which nevertheless can be handled easily using ensemble forecasting. We showed that shallow neural networks, such as NARX, should not be neglected in comparison to DL-techniques; however, LSTMs and CNNs might perform substantially better with a larger data set, where DL really can demonstrate its strengths, which is rarely available in the groundwater domain though.

REFERENCES

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/f6n4g3 (https://doi.org/10/f6n4g3), 2014.

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), Meteorol. Z., p. 22, https://doi.org/10/f5gf49 (https://doi.org/10/f5gf49), 2013.