Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2020-552-RC2, 2020

Answer to Referee #2 regarding the interactive comment on "Groundwater Level Forecasting with Artificial Neural Networks: A Comparison of LSTM, CNN and NARX" by Andreas Wunsch et al.

Received and published: 24 December 2020

We thank Reviewer #2 for the useful suggestions and comments. We are glad to read the overall positive judgement and we think that the suggestions really helped to further improve the manuscript. We answer each of the comments in red.

We have already revised the manuscript, because the discussion is still open, we will upload the revised pdf in a few days.

The authors perform a comparison between NARX, CNN and LSTM on seq to val and seq to seq mode, over a set of wells. The authors investigate not only the performance of the models but also the computational effort required to calibrate them and the effect of the training length which are interesting and useful aspects. Another interesting and novel aspect is the combined approach to hyper parameters tuning and variable selection. The work is well written and exhaustive, making the work reproducible. The set of experiments was properly designed and explained. Therefore, I recommend only a few minor changes.
Thank you.

It could be interesting and useful to show a map of the study area with the well's locations and the mentioned surface water bodies.
We have added a map of the study area to the text including the position of the wells and the available time series lengths for each well. (see Fig3)

In the introduction it could be useful to point out as a novelty aspect the approach to hyper parameters tuning and variables selection.
We have added a statement to the introduction. (See lines 66f.)

Several other works use statistical methods to determine the input sequence length, which could have a hydrological meaning (Kisi et al. 2017; Hasda et al. 2020; Zanotti et al. 2019; Di Nunno 2020). In this case results (Tables S2 – S4) show a wide variability of the length of the input. This does not give any insight about the hydrological behaviour of the water bodies, but it could be useful in cases where the correlation is not linear.
We are aware of using statistical methods for this purpose such as cross correlation between precipitation and groundwater levels. We even performed this ourselves in an earlier study (Wunsch et al. 2018). However, we observed that optimizing the sequence length according to the needs of the network, instead of choosing a length with hydrological meaning, can yield better results. Obtaining the best possible results was definitely the overarching goal here. In fact, one should rather try to interpret the length chosen by the optimization algorithm. However, we did not find any systematics here (for example most models even choose different sequence lengths for same wells, even same models choose different sequence lengths for different setups (seq2val / seq2seq). Additionally, we rely on a poor data basis regarding local geological information for each of the wells, which makes an analysis difficult.

In paragraph 2.5 could be better explained the relationship between the input delay, feedback delay and the additional GWt-1 data. Since NARX is autoregressive isn't it already considering previous groundwater levels? In table S2 you have ID GWLt-1: does it mean that you are feeding into the model more than one GWL (and the same for seq length when using GWLt-1)?
Yes, NARX is autoregressive, hence it feeds back the simulated groundwater level for a certain number of timesteps (defined by the size of the feedback delay). However, we can also provide the NARX simultaneously with a certain number (defined by the size of the input delay) of observed groundwater levels up to t-1 (which is the last known at time t). This way, the NARX model learns the strong relation between the past observed GWLs and the desired output (GWL(t)). This is more exact than simply feeding simulated GWLs back to the model, which have a certain error that might be large. NARX models also have a mode that is known as "open loop" or "series parallel", which follows the same idea of feeding observed values as inputs. However, in this study we do NOT replace the feedback connection as it would be the case for an open loop configuration. This applies both for seq2val and seq2seq scenarios. In the latter case of course, there is an overlap of the sequences that are simultaneously fed to the

network, however, this is the case for all parameters not only past GWLs. We included additional explanations in Sections 2.2 (instead of 2.5), because we thought the they would fit there better.
Lines 106-110., Section 2.2

The performance of the models is well presented and discussed, but a discussion could be made also relatively to very poor performance on some wells: what could be the cause of the results of e.g. BW 781-304-2 or BW 138-019-0? Maybe it could be handy to add the length of the training set in figures S1-S68 or in their captions.
You are right, this aspect deserves more discussion. We added some explanations for other wells that might be useful to understand why forecasting is more challenging and why the performance declines (Lines 303ff). We also agree that additional information on the time series length would be useful. We therefore included an additional figure (Fig.3), which visualizes the available data records for each well. We think this might be handier than giving the length for each time series in the caption.

Fig. 4 and its relative discussion are in the results; it could be useful to mention in the materials and methods that you performed that analysis.
Thank you for this suggestion, however, we kindly disagree. This was really only a side aspect and we did not systematically try to find additional input parameters to improve forecasting performance. We think that it is a nice explanation and visualization within the results section, but we do not want to make it part of the overall experimental setup described in the methods section, which would give it more weight than it deserves in the manuscript. We would feel obligated to perform a proper analysis regarding other influencing parameters, which we could not do, due to lacking data.

Same for paragraph 4.4: it is an interesting analysis, and its methodology should be appropriately explained in the methods section and anticipated in the introduction.
Thank you for this comment, indeed this got somehow lost in the previous version of our manuscript. We added an explanation on our experiments and established a new section on data dependency (Section 2.6, Lines 211ff.) Please compare also comments of Reviewer #1.

Line 50-51 is not clear
We removed the respecting sentence because it gives no added value to the manuscript anyway.

Di Nunno, F., Granata, F., 2020. Groundwater level prediction in Apulia region (Southern Italy) using NARX neural network. Environ. Res. https://doi.org/10.1016/j.envres.2020.110062

Hasda, R., Rahaman, M.F., Jahan, C.S., Molla, K.I., Mazumder, Q.H., 2020. Climatic data analysis for groundwater level simulation in drought prone Barind Tract, Bangladesh: Modelling approach using artificial neural network. Groundw. Sustain. Dev. https://doi.org/10.1016/j.gsd.2020.100361

Kisi, O., Alizamir, M., Zounemat-Kermani, M., 2017. Modeling groundwater fluctuations by three different evolutionary neural network techniques using hydroclimatic data. Nat. hazards 87, 367–381.

Zanotti, C., Rotiroti, M., Sterlacchini, S., Cappellini, G., Fumagalli, L., Stefania, G.A., Nannucci, M.S., Leoni, B., Bonomi, T., 2019. Choosing between linear and nonlinear models and avoiding overfitting for short and long term groundwater level forecasting in a linear system. J. Hydrol. 578, 124015.

Wunsch, A., Liesch, T. and Broda, S.: Forecasting groundwater levels using nonlinear autoregressive networks with exogenous input (NARX), Journal of Hydrology, 567, 743–758, https://doi.org/10/gcx5k4, 2018.