Reply on RC1

Marvin Höge et al.

Author comment on "Improving hydrologic models for predictions and process understanding using neural ODEs" by Marvin Höge et al., Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2022-56-AC2, 2022

We thank the referee for the thorough review, detailed feedback and valuable comments. Please find our replies corresponding to the enumerated points.

Specific comments

- We appreciate the specific recommendation to broaden the coverage of related approaches like PINNs etc. We will add respective information and references to further support the mentioned claim.
- Thank you for raising these points in the discussion. We assume that the referee refers to l. 40ff in the manuscript. We agree, that modelling success is always subject to task, data, system, etc. Yet, for predictive tasks at daily resolution, LSTMs have frequently proven to show overall best results. To address the remark, we suggest, to highlight the "predictive" aspect more strongly in the revision. We do think that Gauch et al. 2021 ("Rainfall–Runoff Prediction at Multiple Timescales with a Single Long Short-Term Memory Network") is the appropriate reference on temporal resolution of LSTM predictions since their work introduces Multi-Timescale LSTMs for this purpose. We think the issue on using different data sources is addressed in Kratzert, F., Klotz, D., Hochreiter, S., and Nearing, G. S.: A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modeling, Hydrology and Earth System Sciences, 25, 2685–2703, 2021. that we have also in our list of references. If we misunderstand the referees comment here, we appreciate further remarks in this respect.
- In the M50 approach, two small, process-specific NNs are used to substitute two different processes: one for discharge Q and the other one for evapotranspiration ET. Qualitatively, temperature T should not have a direct impact on Q but indirectly on the entire water balance via ET (and of course melting M but M is not substituted by an NN in M50). We considered only precipitation P and water storage S_water to have an effect on Q and therefore these are the only input variables to NN^{50}_Q – this is how we show it in Figure 1 (b). Yet, the referee's remark made us aware of the fact that in the top of Figure 1 (b), P is depicted as input variable to the other neural network NN^{50}_{ET}. This should rather be temperature T as it is also used in the mechanistic formulation of ET in Equation A3 and as it is discussed in the manuscript. We thank the referee for spotting typo and we will change the figure accordingly.
- Thank you. For keeping the main article concise, we suggest to add such a requested more detailed figure to the appendix and reference it in the main article.
- *Part 1)* We did not take measures to prevent overfitting during pre-training. We did not

see a need for any sort of regularization in the pre-training phase since this step is only supposed to "roughly" inform the NN parameters such that the NN learns how to process the input variables appropriately to approximate the internal model processes. The target values in this step stem from hard-coded processes relations from the original conceptual and are themselves only approximations to what actually happens in reality. We minimized the sum of squared residuals to these target values as objective function over 1000 iterations in this pre-training phase. *Part 2)* We are happy to provide more details w.r.t. the training: The loss function for the training of the whole Neural ODE models (conceptual model frame + NN(s) within) is NSE. The training is conducted using the GalacticOptim.jl package in Julia where the Neural ODE model parameters (i.e., the parameters of the NN(s) within the conceptual model frame) are optimized using ADAM(): For M100 the only model parameters in the model are the weights and biases of the NN^{100}. For M50, the model parameters that are optimized are the weights and biases of the NN^{50}_Q and NN^{50}_ET (the model parameters in the process $M(t)$ are those from the conceptual model and are kept fixed.) Apart from this optimization, there is no additional fine-tuning. Overall we think that the release of the code will help also with this point (please see our reply to point 12)

- Thank you. We also found Jiang et al., 2020 to serve as a good reference point for a comparison.
- Thank you, this is a very important point. We agree that all readers should be able to easily read the shown figures. Since the major purpose of these figures was to show the ability of the models to match the temporal pattern of hydrographs, snow water equivalent and (model-based) water storage rather than the exact fitting of data-points, we decided that gaps in dashed and dotted lines do not hinder this insight. Yet, we will reiterate the figures and search for suitable alternatives that might serve all mentioned objectives.
- This important point raised by the referee relates to comment 7 above. We chose the color-coding "thermal" having in mind that it is an often recommended set of colors. Yet, now we read "The misuse of color in science communication" by Crameri et al., 2020, Nature Communications, to obtain better knowledge in this field. There, "thermal" is highlighted as one of the colorblindness-friendly options confirming our choice.
- We think that Figure 4 clearly shows that for water storage values smaller than the mean estimated S1, the discharge values estimated by the three different models are very similar while for larger storage values, there is a large discrepancy. This was the main point that we discussed also in the corresponding text in l.287ff. We did not see additional insights in this respect in log-scale plots but we will look into whether providing log-scale plots (maybe in the appendix) might add additional value.
- Thank you for the discussion of this aspect. We agree that extrapolation of neural networks is a challenging topic. In our application case we purposely also picked basins to show that this extrapolation is difficult and sometimes leads to counter-intuitive behavior – to highlight these issues. We still think that more data might help because, first, they help the network to refine relations for data ranges that are closer to the ranges of pure extrapolation. Second, since we consider natural systems where physical limits constrain the problem to be learned by the NN. These are enforced by the chosen model structure and act as regularization aside of the information provided by data (please see our answer w.r.t. regularization for remark 5 above). We think that this combination might help to elicit extrapolation relations that can then be evaluated in plausibility testing. An example for this would be a very high rainfall intensity where our data might never cover a centennial occurrence, but we would still be able to qualitatively judge whether the relation learned by network is plausible or not. Nonetheless, we agree, more data is not necessarily the solution to the problem alone. An extended discussion of this topic is useful and we will elaborate on it.
- We partially agree. There can be confusion if "significance" is used with respect to results from statistical tests, but we do not think that this is the case here. Significant is a frequently used term also outside of only statistical tests and this was our intention

to use it. Otherwise, we would have used "statistically significant". Nonetheless, we will check the different instances to decide whether a substitution by "considerably" or comparable alternatives is an option.

- We fully agree and we will make the corresponding code publicly available. This point was also raised in the community discussion, please see our answer there: "it was always our intention to provide the code. For this purpose, we have set up a Github repository at the time of our submission. However, we wanted to have a discussion about the approach at first in a review process, and only want to release the code of a method that is considered interesting by the scientific community. Now, after receiving encouraging feedback, we will release the code in the coming revision."

Technical corrections

- Thank you for spotting this. As we correctly showed in Equation 3, the text should of course say "observed average" – we will correct this.
- We agree that directly showing and using the equation for NSE is very common in hydrology. Yet, we purposely decided to focus on CoE_alpha and to show that NSE (like the also used mNSE) is only a special case of CoE_alpha. This is not often highlighted in hydrology but we think it is an useful relation to know. We think that lines 203-213 explain this sufficiently and hope that the referee can follow our intention here.
- Thank you, we will correct the typo.