Hydrology and Earth System Sciences Discussions

Hydrol. Earth Syst. Sci. Discuss., referee comment RC1
https://doi.org/10.5194/hess-2022-56-RC1, 2022


# Comment on hess-2022-56

Andreas Wunsch (Referee)

---

Referee 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-RC1, 2022

---

## General Comments

This study uses Neural ODEs to address the gap between process-based models (PBMs) and artificial neural networks/deep learning, a hot topic in current research efforts in hydrology. While the former allows physical interpretation and fosters process/domain knowledge, the latter often shows higher simulation accuracy but lacks the aforementioned interpretability. This is an interesting and well performed study where the authors show the potential of Neural ODEs to combine both domains while simultaneously improving the simulation accuracy. The manuscript is well written and well structured. I like especially the structure of the introduction which is divided into three parts, presenting the state of the art of the major domains (PBMs and Neural Networks) and how they can be brought together by Neural ODEs. The study in general addresses a relevant scientific question within the scope of HESS.

Some weaknesses (in my opinion) are addressed in the following.

## Specific Comments

- While I like the general structure of the introduction, it sometimes lacks the bigger picture. While it is not necessary to elaborate on the bigger picture at the very beginning (as on modeling in hydrology in general, for example), the whole chapter will in my opinion profit from shedding light on other approaches that aim to combine PBMs and DL approaches such as: PGDL (physically guided deep learning) /PIML (physically

informed machine learning) and PINNs. Especially the latter might be worth a look, because they also target differential equations. This should help to really justify the claim in L90: "Yet, none of the pure or novel hybrid machine learning approaches has addressed all the above gaps regarding interpretability, physics and knowledge at once."

- I disagree with the statement that LSTMs generally work well on daily timescales. In my opinion this depends on the modeling task, the data, the system etc. So, while it might be true that peak flows are not captured by LSTM models in some basins, I think with higher temporal resolution this might change to a certain extent. Also, the work by Gauch et al. 2021 is rather an approach to use different data sources than to overcome some model shortcoming. The authors should justify their claim by citing examples or by elaborating why that is in their opinion.
- I am wondering why NN^50_Q did not receive T as an additional input. That would be an obvious choice for a better estimation of ET. Maybe the authors could better elaborate why the chose additional inputs for some of the NNs (such as P to infer runoff) and not for others.
- I think figure 1 is good in general. However, the authors might think of combining the scheme of subplot a with clear indications for (b) and (c), where, e.g., to better explain, which paths/arrows are replaced by which NN model.
- L190 ff. (just as an example): I am missing some details on how the training exactly works.

  First, did you take measures to prevent overfitting while pre-training the NN models? What did you do exactly, if not, please explain why.

  Second, after reading I am still quite unsure how the Neural ODE is trained as a whole. Are the NNs fine-tuned in this step? If yes, how? What is, for example, the Loss function, is it a multi-step procedure which separates the NN training from the overall calibration or is it somehow combined? I appreciated the appendix in general, however, in this regard it did not help me to answer my questions.
- I generally like the stringent way of comparing with the results from Jiang et al. 2020.
- The authors should improve the readability of figure 3. Where data points are plotted, I hardly recognize the rest of the plot. Maybe another color-concept might help in general, also to avoid dotted and dashed lines, which are hard to read in my opinion (because the gap between dashs and dots is almost larger than some of the peaks). I generally recommend to select different colors that separate well from each other, and use slight transparency to ensure readability in overlapping regions. In my opinion, being aware of colorblind readers is more useful than ensuring readability when printed/or when black&white. Maybe check colorbrewer or related sites for ideas.
- You might also check if the colormaps in figures 5 to 7 could be improved in terms of readability for colorblind people ( I suspect this is already okay – not sure)
- L287: are they that similar? Probably, but due to the scale it is hard to recognize. Did you explore also log-scale plots? Did this provide additional insights in this regard?
- Figures 5 to 7: As you discuss yourself, what you do is using neural networks for strong extrapolation. Of course this is an application, where ANNs fail on a regular basis. So while I think that this is not necessarily the case, we simply do not know. I would recommend to better discuss this aspect. In L. 395, you state that "20 years of data is not enough to extrapolate towards these limits". I would counter that even with more data this is not necessarily possible. Please discuss the general problems with using ANNs/DL for extrapolation and why you might think that this applies (or not) for your application and analysis.
- Please revise your wording when speaking of significance. Lines: 317, 328, 345, 358, 360 … (and others). If you did perform a statistical test please add the information, otherwise, use e.g., "considerably".
- The software code should be made available. This might have helped to answer some of my questions and will foster the application of this approach in the future.

**Technical Corrections**

- Statement in L. 204-205: "With NSE < 0, the model is worse than just using the model 205 average, […]" – NSE compares to the observed average, not to the model average.
- I find it quite unintuitive to mix NSE, mNSE and CoE_alpha. NSE is more common in hydrology, why not only name it consistently like this?
- L331 spelling: evapotranspiration