Editor Comments: Alexander Gruber

Comments/Text of Referee posted in **black**, our text in **purple**.

We thank the editor for confirming the important aspects of the reviewers comments. We have addressed the following comments and believe that the manuscript is definitely improved!

Dear authors,

thank you for your submission. The referees are, overall, very positive about your manuscript and I share their opinion that the presented study provides an interesting and valuable novel approach to tackle the problem of interpretability in ML for Earth science problems.

I thus invite you to carefully revise your manuscript by implementing your proposed changes to the manuscript, which I believe do, in general, address most of their concerns. I, personally, share the following concerns of the reviewers in particular:

(i) Referee #1: I agree that it might be difficult for readers not too familiar with LSTMs and some other concepts (such as elastic nets) to understand all that is presented. While I also acknowledge that some background is provided and that it is difficult to strike a good balance, I tend to **think that a little more detail (perhaps as am illustrative figure) might be beneficial for a broader readership.**

We have added an Appendix Section F and Figure F1 which give more detail on the LSTM, as well as providing a list of more complete references for readers who may want to further explore the LSTM structure.

(ii) Referee #2: I share the opinion that statements such as "the LSTM produces more accurate discharge simulations than any other hydrological model" need to be toned down a bit. In my view, there isn't a large enough body of evidence that would proof that LSTMs universally perform better than hydrological models, which I don't believe is what the authors want to say, but the text somewhat reads as so…

We agree and have changed our statements as outlined in the responses to Reviewer 2.

L278 "*Firstly, previous studies have demonstrated that the LSTM produces more accurate discharge simulations when benchmarked against other hydrological models, and so we might expect that the intermediate variables are also better represented by the LSTM.*"

L292 "*Since the LSTM was found to be the most accurate rainfall-runoff model for discharge in a series of studies (Kratzert et al., 2018, 2019c; Gauch et al., 2021; Frame et al., 2021; Gauch et al., 2020) it makes sense to explore the soil moisture that the LSTM associates with a simulated level of discharge.*"

Would it perhaps be more fair to say that physically based models should - in theory - be more generalisable (by definition), yet they often require model calibration to account for representativeness errors; whereas ML-based approaches learn whatever data you throw at them and MAY be generalisable provided that actual physical processes are learned? I believe this study makes a good case that it could, and I believe this is also its intention... The fact that a well-trained non-linear ML model might, in the presence of good training data, often - if not always - outperform a perhaps not perfectly calibrated physical model with various simplifying assumptions etc. solely in terms of predictive accuracy is, from how I see it, not really a relevant statement in any context.

Thank you for the effective summary of one of our conclusions. In order to more concretely address this point we have added the following statement in the conclusion:

"*This finding offers a potential explanation for recent research that showed LSTMs have the potential to generalise to out of sample conditions \citep{frame2021deep}, since they have learned physically realistic mappings from inputs to outputs.*"

Best regards,
Alexander Gruber